WIDE-BAND BUTTERFLY NETWORK: STABLE AND EFFICIENT INVERSION VIA MULTI-FREQUENCY NEURAL NETWORKS.*

MATTHEW LI†, LAURENT DEMANET‡, AND LEONARDO ZEPEDA-NÚÑEZ§

Abstract.
We introduce an end-to-end deep learning architecture called the *wide-band butterfly network* (WideBNet) for approximating the inverse scattering map from wide-band scattering data. This architecture incorporates tools from computational harmonic analysis, such as the butterfly factorization, and traditional multi-scale methods, such as the Cooley-Tukey FFT algorithm, to drastically reduce the number of trainable parameters to match the inherent complexity of the problem. As a result, WideBNet is efficient: it requires fewer training points than off-the-shelf architectures, and has stable training dynamics which are compatible with standard weight initialization strategies. The architecture automatically adapts to the dimensions of the data with only a few hyper-parameters that the user must specify. WideBNet is able to produce images that are competitive with optimization-based approaches, but at a fraction of the cost, and we also demonstrate numerically that it learns to super-resolve scatterers with a full aperture configuration.

1. Introduction. There is nowadays extensive documentation on the remarkable ability of neural networks to approximate high-dimensional, non-linear maps provided that enough data are available [56]. In many applications the process of discovering such approximations simply involves enriching the network models, i.e., making them wider and/or deeper, until favourable stationary points arise in the empirical loss landscape. This practice can be partially justified by the asymptotic capacity of neural networks to approximate functions to within arbitrary accuracy, assuming only mild regularity conditions [22, 46, 66]. Oftentimes, however, this strategy results in models that are vastly over-parametrized, even when compared to the already massive datasets that are necessary for training. For reasons that we outline below, these approximation-theoretic results also obscure many pre-asymptotic complications that are particularly acute when neural networks are applied to scientific applications. In these instances the neural architectures often require specific tailoring to the task at hand in order to satisfy the stricter requirements of scientific computing.

In this paper we focus on the problem of high-resolution imaging of scatterers arising from wave-based inverse problems. This task naturally arises in many scientific applications: e.g. biomedical imaging [79], synthetic aperture radar [20], non-destructive testing [72], and geophysics [76]. This problem also prototypically exhibits two challenges that are commonly encountered in scientific machine learning. First: obtaining the training data in this setting – whether synthetically or experimentally – comes at considerable expense, which bottlenecks the size of the models that can be reliably trained to satisfy the stringent accuracy requirements. This necessitates the use of unconventional architectures that are bespoke to each problem. Second: wave scattering involves *non-smooth data* that are recordings of highly oscillatory, broad-

*Submitted to the editors DATE.
Funding: The authors thank Total SA for support. LD is also supported by AFOSR grant FA9550-17-1-0316. L.Z.-N. is supported in part by the Wisconsin Alumni Research Fund, the National Science Foundation under the grant DMS-2012292, and NSF TRIPODS award 1740707.

†Computational Science and Engineering, Massachusetts Institute of Technology, Cambridge MA 02139 (mtcli@mit.edu)
‡Department of Mathematics and Earth Resources Lab, Massachusetts Institute of Technology, Cambridge MA 02139 (laurent@math.mit.edu)
§Department of Mathematics, University of Wisconsin-Madison, Madison WI 53706 (zepeda-nunez@wisc.edu)

band, scattered waveforms. These highly oscillatory (i.e. high-frequency) signals are known to impede the training dynamics of many machine learning algorithms [88] and thus require new strategies to mitigate their effect.

Existing methods for scientific machine learning address the issue of data scarcity by "incorporating underlying physics" into the design of neural architectures. In instances where the problem data are *smooth*, this demonstrably reduces the total number of trainable weights which, in turn, reduces the number of training data required. Broadly categorized, these designs manifest as either: *(i)* explicitly enforcing physical symmetries into the network [92, 95, 96], *(ii)* exploiting signal invariances and equivariances when processing the data [12], *(iii)* directly embedding the governing differential equations into the objective function [48, 75], or *(iv)* imposing information flow (i.e., connectivity) within the architecture according to multi-scale interactions inherent to the physics of the data generating process [34, 51]. Surprisingly, in addition to lowering data requirements these strategies are also observed to improve on the testing accuracy of comparable conventional models which are trained on a larger set of training points [44, 66, 94].

In comparison, not much is known about designing architectures for processing *non-smooth data* such as high-frequency waves. Here the same challenge that confounds the original inverse problem – namely, the processing of highly oscillatory signals – similarly obstructs direct application of machine learning methods. This idea is formalized by the "F-principle" conjecture [88] which documents the relation between machine learning methods and Fourier analysis. Specifically, it is empirically observed that models with fully-connected and convolutional architectures preferentially capture the low-frequency features of the target function. On the other hand, considerable expense (with respect to model size and/or data) is needed to learn high-frequency features [64]. Some examples even demonstrate that training can completely fail when the target function lacks low-frequency content even if highly expressive models are used [15, 87]. The F-principle thus demonstrates that although neural networks are universal approximators in an asymptotic sense, new strategies are needed to account for the issue with high frequencies if tractably computable models are to be obtained.

We note that in our application the forward and inverse maps are intrinsically oscillatory on account of the physics of wave propagation. This can be seen as an immediate consequence of the *dispersion relation* in homogeneous media,

$$\lambda f = c, \tag{1.1}$$

which describes the inverse scaling of the frequency f of propagating waves to their spatial wavelength λ by a factor of the local wave-speed c. This dispersion relation, in conjunction with rudimentary signal processing, effectively suggests that images generated by back-propagating the recorded waves into the medium are constrained to a wavelength dependent resolution limit, i.e., the classical diffraction limit [39]. High resolution imaging of scatterers thus seemingly necessitates the use of high frequency waves to probe the media.

1.1. Our Contributions. We introduce a custom architecture for the inverse wave scattering problem which we call WideBNet. We demonstrate that our architecture overcomes the major deficiencies outlined above for traditional architectures. Specifically, WideBNet relies on ideas from the butterfly factorization [59] to capture the Fourier Integral Operators (FIOs) underlying the physics of wave-scattering – as a result, fewer training datapoints are needed. Moreover, it addresses the high frequency limitations identified by F-principle by mimicking the Cooley-Tukey algorithm [27] to

process multi-frequency data only at localized length scales – this effectively renders each frequency slice as *locally* low-frequency information. These design choices afford WideBNet the following benefits compared to off-the-shelf deep learning models:

Training Efficiency The architecture builds upon the butterfly factorization and thus systematically adapts to the input size of the data, i.e., the number of pixels in the image. As a result, the degrees of freedom in the model scale near-linearly with the input size, and the depth of the network scales logarithmically with the input size[1]. This makes training our network data-efficient as there are relatively fewer degrees of freedom.

Training Stability WideBNet avoids empirically observed shortcomings with other network architectures that rely on the butterfly factorization. For example, in [58] the authors prove that butterfly-networks are capable of efficiently approximating generic FIOs, but report that learning such operators requires an accurate initialization to avoid local minima; this is typically not easily obtainable for most FIOs, including our application. Similarly, [51] introduces a butterfly-network for single frequency inversion but requires increasing the width of their network (so that the degrees of freedom no longer scale linearly) to overcome local minima. In contrast, empirically we observe that WideBNet does not require specialized initialization strategies, it does not routinely get stuck in local minima, and it does not exhibit exploding or vanishing gradients. We speculate that the training stability of WideBNet can be attributed to its use of multi-frequency data that are banded to appropriate length scales to avoid the F-principle limitations.

Imaging Super-resolution In our numerical results we demonstrate that our network super-resolves scatterers, i.e., produces sharp images of sub-wavelength features[2] such as diffraction corners, in addition to producing competitive images when compared against classical optimization-based inversion methods in the traditional super-diffraction regime.

Hyper-parameter Efficiency It is efficient to tune the hyper-parameters of WideBNet as there are only a few which are used to describe the architecture. We note that in numerical examples we observe strong robustness to variations in these hyper-parameters. This indicates that relatively little effort is needed on the user's part to optimally tune our architecture.

A detailed discussion of the WideBNet architecture, as well as implementation notes, can be found in Section 3. Meanwhile, we briefly sketch the intuition behind the design choice here. The idea to embed the butterfly factorization into the architecture is to effectively furnish our network with a strong prior on the physics of wave scattering. Indeed, we provide numerical evidence that it is necessary to manually encode the long range "non-local" interactions between scatterers and sources that are inherent to the wave kernel. Mathematically these interactions are known

[1] When compared to other machine learning based approaches, we note that a comparable implementation using fully connected networks results in models with degrees of freedom that scale cubically with the size of the input, i.e., the number of pixels in the image, and are thus prohibitively expensive to train. Conversely, a purely convolutional neural network implementation for the task requires far deeper networks (or far wider filters) to properly capture the long-range interactions governed by the underlying wave physics. Such deep networks are known to exhibit issues with exploding/vanishing gradients leading to unstable training dynamics [7]. While we do not discount the possibility of other hybridized (fully connected + convolutional) architectures which achieve the same task, we emphasize that these architectures would not be immediately transferable for different image and data resolution requirements.

[2] We plan to further investigate and document this super-resolution phenomenon in forthcoming work.

to described as the action of an FIO [45], which can be discretely represented in a complexity-optimal manner by means of the butterfly factorization [59] and the butterfly algorithm [16, 70, 11].

However we stress that the marriage of the butterfly factorization with network architectures is *not* the original contribution of this work; butterfly-like architectures have been previously proposed by other authors, albeit with different goals [58, 51], and we review these contributions below in Section 1.2. Instead, our contribution is the *combination of this network architecture with multi-frequency data*. This data assimilation strategy takes cues from the Cooley-Tukey algorithm and is done, in part, to address the F-principle. For reference, one notable strategy for avoiding the F-principle involves partitioning the model into disjoint Fourier segments and frequency down-shifting accordingly [14], but this introduces costly convolutions in data-space and requires a dense data sampling strategy that scales unfavourably with dimensionality. Our network improves on this approach by exploiting the duality between frequency f and wavelength λ, as described by the dispersion relation in (1.1), to introduce data only at their local length scales. This effectively performs frequency downshifting by *spatial downsampling*. This strategy is easily accommodated by the butterfly architecture as these multi-scale interactions are already implicitly present in its formulation.

Outline. The remainder of this document is structured as follows. We close this section with relevant background material on existing algorithms for inverse scattering and relevant machine learning based approaches for general inverse problems in Section 1.2. Section 2 describes the technical details of the underlying physical model, provides background on the problem to solve and the algorithmic ideas behind the network. In Section 3 we present in detail the network architecture. Finally, in Section 4 we present and discuss the numerical results.

1.2. Related Literature.

1.2.1. Classical Approaches. One of the earliest modalities in imaging is travel-time tomography [69, 43, 5], in which the travel time of a wave passing between two points is used to reconstruct the medium wave-speed [80]. Travel-time tomography is a rather mature technique, which can even be easily and cost effectively implemented in portable ultra-sound devices [23]. However, its resolution deteriorates greatly when dealing with highly heterogeneous media and in the presence of multiple scattering.

In response to these drawbacks, several techniques were developed such as reverse time migration [6], linear sampling method [24], decomposition methods [54] among many others. See [26] and [85] for excellent historical reviews.

Finally, a high-resolution technique, called full-waveform inversion (FWI) [82] was developed in the late 80s, which has been shown empirically capable of handling multiple scattering. FWI solves a constrained optimization problem in which the misfit between the real data and synthetic data coming from the numerical solution of the PDE is minimized. This technique, coupled with large computing power, has been successful at recovering the properties of the sub-surface [74]. Nowadays, it is considered the gold standard in geophysical exploration [84].

Despite its enormous success, FWI still suffers from three significant challenges: prohibitive computational cost, cycle-skipping and limited resolution. The prohibitive computational cost is linked to the cost of computing the gradient within the optimization loop, which requires a large amount of wave solves. The resulting complexity

of each iteration is quadratic[3] [10] with respect to number of unknowns to recover. Progress in this direction has focused on developing fast PDE solvers [93, 33] which are necessary to compute the gradient. In addition, numerous iterations are usually required for convergence. This prohibitive computational cost has hampered the application of this vastly superior technique to domains where images are required on-the-fly, such as biomedical imaging.

Cycle-skipping refers to the undesirable convergence to spurious local minima by the FWI algorithm. This effect is especially pronounced when low-frequency data are scarce as these determine the kinematically relevant, low-wavenumber components of the material properties. Unfortunately, acquiring low-frequency data from practical field applications is a challenging and expensive task. As such, research in this area has focused on regularizing the optimization objective to handle the lack of low-frequency data [81, 83], using a smooth initial guess from travel-time tomography [3], or extrapolating the low-frequency component from higher frequency data [61]. Lastly, quantifying the resolution limits of FWI remains an open problem [38], i.e., understanding the finest details available by the algorithm and its scaling with respect to the shortest wavelength at which data are available. This is important for applications requiring accurate images of discontinuities [4, 13, 30, 29], such as those arising in natural geophysical formations, for properly detecting cracks and dislocations in materials, or for detecting and interpreting anomalies in biomedical imaging.

1.3. Machine Learning Approaches. Besides the classical, PDE constrained optimization approaches, several recent methodologies based on machine learning for more general inverse problems have been proposed lately.

In [19] authors used the recently introduced paradigm of physics informed neural networks (PINN) to solve for inverse problems in optics. Aggarwal et al. introduce a model-based image reconstruction framework [2] for MRI reconstruction. The formulation contains a novel data-consistency step that performs conjugate gradient iterations inside the unrolled algorithm. Gilton et al. proposed in [40] a novel network based on Neumann series coupled with a hand-crafted pre-conditioner for linear inverse problems, which recast an unrolled algorithm as elements of a Neumann series. In [65] Mao et al. use a deep encoder-decoder network reminiscent of U-nets [78] for image de-noising, using symmetric skip connections.

In [35] the authors proposes a rotationally equivariant network for inverse scattering, that is only valid for homogeneous media; the same type of ideas is applied to travel-time tomography [37] and optical tomography [36].

Among the more general field of computational harmonic analysis, to which the butterfly algorithm is connected, we mention several other applications. Networks based on the Short-time Fourier transform [90, 89] has been used for hierarchically decomposing signals in a non-linear fashion. Networks based on the scattering transform has been proposed [12] to take in account translation invariance in images. In [91] the authors introduced another framework based on frames for inverse problems, which was applied to computer tomography de-noising [50].

In addition, machine learning recently has been used for super resolution in the signal processing context [18] and image processing. Recently newly developed frameworks such as generative adversarial networks (GANs) [41, 42], and variational autoencoders (VAEs) [53, 77] have been used for super resolution in the context of image processing [49, 57, 67]. These techniques provide an end-to-end map that relies on

[3]Using state-of-the-art sparse direct solvers. It can be further reduced to $\mathcal{O}(N^{3/2})$ using state-of-the-art pre-conditioner, but with substantially larger constants.

the statistical properties of the images to super-resolve them.

Another related approach is the recently introduced Fourier Neural Operators [62] that aims to learn the Fourier multipliers in a context akin to pseudo-differential operators using an aggressive filtering, which is compensated by the non-linear activation functions. Although this approach captures long-range interactions, it is unclear whether the highly oscillatory behavior of wave data can be captured efficiently.

The method introduced in this manuscript follows similar ideas to [34, 28], where the authors introduce tools from numerical analysis into deep learning. They build on the sparse matrix factorizations that result from exploiting low-rank interactions arising from the underlying physics of the problem. These factorizations are translated into the machine learning context: each matrix factor becomes a layer in the network wherein the sparsity pattern informs the connectivity between layers, and the matrix entries themselves are viewed as learnable weights. In particular, the authors translate hierarchical matrices ($\mathcal{H}$-matrices), which are factorizations of operators into low-rank and permutations matrices, into individual layers in neural network architectures. Although these networks are well suited for smooth data with compressible long range interactions, which is the underlying motivation for the $\mathcal{H}$-matrices, they are not well suited for wave-scattering problems where the data are highly oscillatory, and where the long-range interactions are not typically compressible.

Instead, the correct idea for capturing wave propagation is the choice of the butterfly factorization, as motivated by their use for representing FIOs. In fact, architectures based on butterfly algorithm have been previously proposed, albeit with different goals as the one considered in this paper. In [28] the authors recover the butterfly structure of certain linear operators, from permutation operations. In [51] the authors use a one-level butterfly network with applications to inverse scattering, though critically they require a super-linear scaling in their number of parameters. In [58] the authors propose a mono-chromatic butterfly network similar to the architecture used in this case, which was later simplified in [86]. In [28], the authors use the backbone of the butterfly structure to learn fast matrix approximations, with a clever variational relaxation strategy for learning the permutation factors. However, as mentioned in the prequel, none of these works address the use of butterfly factorizations for super-resolution in wave-based imaging which requires stable training over a wideband dataset.

2. Background.

In this section we briefly review concepts from classical imaging (see [25] for further details) and their connection with fast numerical methods. We also provide a succinct description of the butterfly factorization and Cooley-Tukey FFT algorithm to motivate the discussion of our architecture in Section 3.

2.1. Underlying Physical Model.

We consider the time-harmonic wave equation with constant-density acoustic physics, also called the Helmholtz equation, with frequency ω and squared slowness m, given by

$$(\Delta + \omega^2 m(\mathbf{x}))u(\mathbf{x}) = 0 \tag{2.1}$$

with radiating boundary conditions. We further suppose the slowness squared admits a scale separation into

$$m(\mathbf{x}) = m_0(\mathbf{x}) + \eta(\mathbf{x}),$$

where m_0 corresponds to the smooth background slowness, assumed known, and η the rough perturbation that we wish to recover. If the background slowness is constant

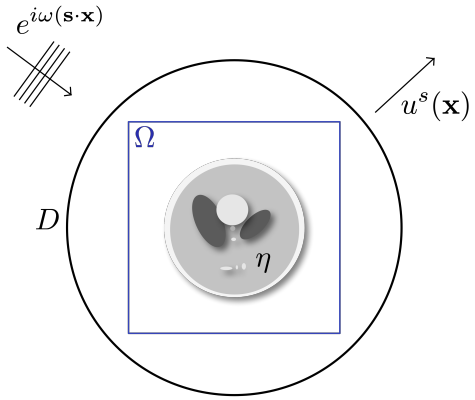

FIG. 1. *Diagram of the inverse scattering problem. We probe the medium with a plane-wave with direction* $\mathbf{s}$*, and we sample the scattered field on the disk* D*.*

and normalized[4] so that

$$m(\mathbf{x}) = 1 + \eta(\mathbf{x}),$$

then solutions to (2.1) can be expressed in the form

$$u(\mathbf{x}) = e^{i\omega(\mathbf{s}\cdot\mathbf{x})} + u^{sc}(\mathbf{x}), \tag{2.2}$$

where $e^{i\omega(\mathbf{s}\cdot\mathbf{x})}$ is the incoming plane wave, with propagating direction $\mathbf{s}$, that we use to "probe" the perturbation, and $u^{sc}(\mathbf{x})$ is the scattered field produced by the interaction of the perturbation with the impinging wave. The scattered field satisfies

$$\begin{cases} \left(\Delta + \omega^2(1+\eta(\mathbf{x}))\right) u^{sc}(\mathbf{x}) = -\omega^2 \eta(\mathbf{x}) e^{i\omega(\mathbf{s}\cdot\mathbf{x})} & \text{for } \mathbf{x} \in \mathbb{R}^2, \\ \lim\limits_{|\mathbf{x}|\to\infty} |\mathbf{x}|^{1/2} \left(\dfrac{\partial}{\partial|\mathbf{x}|} - \mathrm{i}\omega\right) u^s(\mathbf{x}) = 0, \end{cases} \tag{2.3}$$

following the configuration depicted in Fig. 1. We select the detector manifold D to be a circle of radius R that encloses the domain of interest Ω. For each incoming direction $\mathbf{s} \in \mathbb{S}^1$, as defined in (2.3), the data are given by sampling the scattered field with receiver elements that are located on D and indexed by $\mathbf{r} \in \mathbb{S}^1$. We assemble the data for each frequency ω into a matrix Λ^ω whose $(\mathbf{s},\mathbf{r})$-th entry corresponds to

$$\Lambda_{\mathbf{s},\mathbf{r}} = u^{sc}(R\mathbf{r}; \mathbf{s}), \tag{2.4}$$

where we omit the dependence on ω in the right hand side. We call $\mathcal{F}^\omega[\eta]$ the *forward map* relating the perturbation η to the data matrix Λ^ω[5].

Accordingly, we can cast the inverse problem for recovering the rough perturbation as

$$\eta^* = \mathrm{argmin}_\mu \|\mathcal{F}^\omega[\mu] - \Lambda^\omega\|_2^2. \tag{2.5}$$

[4]This assumption is only made to make the presentation more transparent.

[5]We point out that the data is not linearized, we solve (2.3), which depends non-linearly on η, to obtain the scattered wavefield, u^{sc}, for each incoming direction. One can easily recover the full wavefield using (2.2).

Linearizing the forward operator $\mathcal{F}^\omega$ is instructive as it sheds light on the essential difficulties of this problem. Using the classical Born approximation in (2.3) we obtain that

$$u^{sc}(\mathbf{x}) = \omega^2 \int_{\mathbb{R}^2} \Phi^\omega(\mathbf{x},\mathbf{y})\eta(\mathbf{y})e^{i\omega(\mathbf{s}\cdot\mathbf{y})}d\mathbf{y}, \tag{2.6}$$

where Φ^ω is the Green's function of the two-dimensional Helmholtz equation in homogeneous media, i.e., Φ^ω satisfies

$$\begin{cases} \left(\Delta + \omega^2\right)\Phi^\omega(\mathbf{x},\mathbf{y}) = -\delta(\mathbf{x},\mathbf{y}) & \text{for } \mathbf{x}\in\mathbb{R}^2, \\ \lim\limits_{|\mathbf{x}|\to\infty} |\mathbf{x}|^{1/2}\left(\dfrac{\partial}{\partial|\mathbf{x}|} - \mathrm{i}\omega\right)\Phi^\omega(\mathbf{x},\mathbf{y}) = 0. \end{cases} \tag{2.7}$$

Furthermore, we can use the classical far-field asymptotics of the Green's function to express

$$u^{sc}(R\mathbf{r}) = -\omega^2\frac{e^{i\omega R}}{\sqrt{R}}\int_{\mathbb{R}^2}\eta(y)e^{i\omega(\mathbf{s}-\mathbf{r})\cdot\mathbf{y})}d\mathbf{y} + \mathcal{O}(R^{-3/2}). \tag{2.8}$$

Thus, up to a re-scaling and a phase change, the far-field pattern defined in (2.4) can be approximately written as a Fourier transform of the perturbation, viz.,

$$\Lambda_{\mathbf{s},\mathbf{r}}(\omega) \approx F^\omega\eta = -\omega^2\frac{e^{i\omega R}}{\sqrt{R}}\int_{\mathbb{R}^2} e^{i\omega(\mathbf{s}-\mathbf{r})\cdot\mathbf{y}}\eta(y)d\mathbf{y}. \tag{2.9}$$

In this notation F^ω is the linearized forward operator acting on the perturbation.

Solving the inverse problem (2.5) using the linearized operator in (2.9) and Tikhonov-regularization with regularization parameter ϵ results in the explicit solution

$$\eta^* = \left(\left(F^\omega\right)^* F^\omega + \epsilon I\right)^{-1}\left(F^\omega\right)^*\Lambda^\omega. \tag{2.10}$$

This formula is also referred to as filtered back-projection [25], is optimal with respect to the L^2-objective and, concomitantly, tends to yield low-pass filtered estimates, particularly with large ϵ. In practice ϵ is chosen to be sufficient large so as to remedy the ill-conditioning of the normal operator $\left(F^\omega\right)^* F^\omega$.

Performing the inversion numerically requires discretizing the wavespeed and the sampling geometry. We discretize Ω using $N = n_x \times n_z$ degrees of freedom following the Nyquist sampling rate of $n_x \sim n_z \sim \omega$. The scattered data Λ^ω are discretized into an $n_{src} \times n_{rcv}$ matrix.

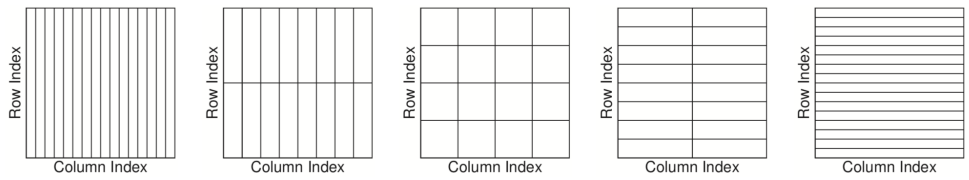

FIG. 2. *Sketch of a matrix exhibiting a complementary low-rank property. Each of the blocks induced by the different partitions has the same ϵ-rank.*

After discretization and a change of variables, $(F^\omega)^*$ in (2.10) is a Fourier transform (which itself is an FIO), and F^*F is a pseudo-differential operator, which, when

the background medium is constant, is translation invariant, thus $(F^*F + \epsilon I)^{-1}$ can be reduced to a convolution-type operator. In more general situations of smooth background media the operator $(F^*F + \epsilon I)^{-1}$ can be approximated by networks specifically tailored for pseudo-differential operators, such as the multiscale-neural network [34].

Remark: Thus far we have assumed that we probe the perturbation η using only a monochromatic time-harmonic wave with fixed frequency ω. As mentioned in the introduction this is known to be ill-posed and data at additional frequencies are required to stabilize the reconstruction [47]. In particular, a time-domain formulation known as the *imaging condition* yields a more stable reconstruction using the full frequency bandwidth; this formula can be formally stated as

$$\eta^* = \int_{\mathbb{R}} \left((F^\omega)^* F^\omega + \epsilon I \right)^{-1} (F^\omega)^* \Lambda^\omega d\alpha(\omega), \tag{2.11}$$

where $d\alpha(\omega)$ is a density related to the frequency content of the probing wavelet. When the density is well approximated by a discrete measure then

$$\eta^* \approx \sum_{i=1}^{N_{\text{freqs}}} \left((F^{\omega_i})^* F^{\omega_i} + \epsilon I \right)^{-1} (F^{\omega_i})^* \Lambda^{\omega_i} \alpha(\omega_i), \tag{2.12}$$

over a discrete set of frequencies $\{\omega_i\}_{i=1}^{N_{\text{freqs}}}$. We note that the selection of these frequencies, in addition to the optimal ordering in which the summation is computed under an iterative regime, remains an open question and an area of active research [10].

2.2. Butterfly Factorization and Fourier Integral Operator. When the scattered field is given by (2.9) then one could apply the fast Fourier transform [27] to compute the estimate (2.12) in quasi-linear time. However, with a heterogeneous background the linearized forward map is instead given by a more general representation usually known as a Fourier integral operator (FIO), which has the form

$$(F^\omega \eta)(\mathbf{x}) = \int_{\mathbb{R}^2} a_\omega(\mathbf{x}, \mathbf{y}) e^{i\omega\phi(\mathbf{x},\mathbf{y})} \eta(\mathbf{y}) d\mathbf{y}. \tag{2.13}$$

Here $\phi(\mathbf{x}, \mathbf{y})$ is referred to as the phase (or travel-time) function while a_ω is typically a very smooth function that encodes the amplitude[6]. The work of [16, 73, 70] recognized that even in this more generalized instance the application of F^ω and its adjoint can be computed in optimal complexity by means of the butterfly algorithm. The butterfly algorithm is a multi-scale algorithm which takes advantage of the *complementary low-rank property* of the discretized operator depicted in Fig. 2. In its original form the algorithm relies on explicit knowledge of the phase function; later, in [59] the authors introduced the butterfly factorization, which approximates the discretized operator (2.13) by the multiplication of sparse matrices with a *specific* sparsity pattern[7] as shown in Fig. 3.

In a nutshell, the butterfly factorization approximately factorizes a matrix A that satisfies the complementary low-rank property in $L + 3$ sparse matrices following:

$$A \approx A_{butterfly} = U^L G^{L-1} \cdots G^{L/2} S^{L/2} \left(H^{L/2} \right)^* \cdots \left(H^{L-1} \right)^* \left(V^L \right)^*, \tag{2.14}$$

[6]The principal symbol a_ω depends asymptotically on ω as $\mathcal{O}(\omega^{-1})$ [68].

[7]This pattern is for the one-dimensional butterfly factorization, which already captures the key algorithmic ideas while keeping the presentation clean of ordering issues that arises in higher dimension.

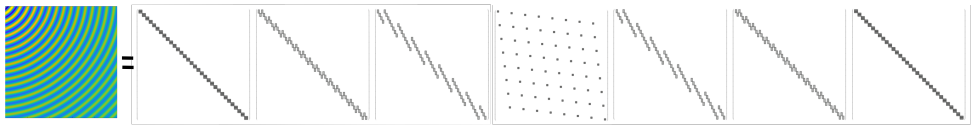

FIG. 3. *Sketch of the butterfly factorization, where the matrix at the left is factorized in sequence of very sparse matrices with a distinct sparsity pattern, induced by Fig. 2.*

where U^L and V^L are block diagonal matrices, $S^{L/2}$ is a weighted permutation matrix, usually called a *switch matrix*, and L is the number of levels in the factorization, which is usually a power of two.

We can interpret the factors in (2.14) following the original butterfly algorithm. V^L extracts a local representation of the vector, then each factor H^ℓ compresses two neighboring local representations, i.e., decimates by a factor of two the number of local representations, while increasing the amount of information in each presentation. The switch matrix $S^{L/2}$ quickly redistributes the information contained in each local representation. The factors G^ℓ decompress the information contained in each representation at each stage, i.e., the local representations are split in two by each factor increasing the spatial resolution, and finally the factor U^L, transforms the local representations to the sampling points.

For the sake of completeness we provide a formal argument to show that the FIO in (2.13) satisfies the complementary rank property (see [16] for a more comprehensive argument). In a nutshell, the complementary rank property for a matrix is the property in which each block of the partition in Fig. 2 have ϵ-ranks bounded by the same constant. Equivalently, any block in which the multiplication of its sides is equal to $\mathcal{O}(N)$ has a bounded ϵ-rank.

Suppose that we have two points $\mathbf{x}_0$ and $\mathbf{y}_0$ in the evaluation and integration region respectively. We define two neighborhoods around each point, such that $|\mathbf{x} - \mathbf{x}_0| < d_x$ and $|\mathbf{y} - \mathbf{y}_0| < d_y$. In this case, d_x and d_y are the sides of the blocks, in physical space, shown in Fig. 2. We then seek to find the largest values of d_x and d_y such that we can efficiently approximate

$$\int_{|\mathbf{y}-\mathbf{y}_0|<d_y} a(\mathbf{x},\mathbf{y}) e^{i\omega\phi(\mathbf{x},\mathbf{y})}\eta(\mathbf{y})d\mathbf{y}, \tag{2.15}$$

using a separable function. The principal symbol, $a(\mathbf{x},\mathbf{y})$ is supposed to be smooth and independent of ω (or weakly dependent), so we can focus our discussion to the oscillatory term $e^{i\omega\phi(\mathbf{x},\mathbf{y})}$.

Using a Taylor expansion we have that

$$\begin{aligned}\phi(\mathbf{x},\mathbf{y}) = &\phi(\mathbf{x}_0,\mathbf{y}_0) + \partial_{\mathbf{x}}\phi(\mathbf{x}_0,\mathbf{y}_0)\cdot(\mathbf{x}-\mathbf{x}_0) + \partial_{\mathbf{y}}\phi(\mathbf{x}_0,\mathbf{y}_0)\cdot(\mathbf{y}-\mathbf{y}_0)\\ &+(\mathbf{x}-\mathbf{x}_0)^T\cdot\partial^2_{\mathbf{x}}\phi(\mathbf{x}_0,\mathbf{y}_0)\cdot(\mathbf{x}-\mathbf{x}_0) + (\mathbf{y}-\mathbf{y}_0)^T\cdot\partial^2_{\mathbf{y}}\phi(\mathbf{y}_0,\mathbf{y}_0)\cdot(\mathbf{y}-\mathbf{y}_0)\\ &+2(\mathbf{x}-\mathbf{x}_0)^T\cdot\partial^2_{\mathbf{x},\mathbf{y}}\phi(\mathbf{x}_0,\mathbf{y}_0)\cdot(\mathbf{y}-\mathbf{y}_0) + \mathcal{O}(d_x d_y)\end{aligned}$$

Clearly the first five terms provide separable expressions, the sixth term can be easily bounded producing

$$e^{i\omega\phi(x,y)} = e^{i\omega\psi(x)}e^{i\omega\xi(y)}(1+\mathcal{O}(\omega d_x d_y)) \tag{2.16}$$

thus as long as $d_x d_y \leq \omega^{-1}$, then $e^{i\omega\phi(x,y)}$ can be locally approximated by a separable function. In the discrete case this property is translated to the fact that the multiplication of the height and the width of each block has a constant ϵ-rank, which is exactly the complementary low-rank property showcased in Fig. 2.

Remark: We point out that there exist three different types of butterfly factorizations. The left one-sided, the right one-sided, and the two-sided (see [63] for a review). In this work we focus on the two-sided version, which provides the best complexity. It is possible to "neuralize" the other two types of factorizations, which yield a specific type of CNN networks with sparse channel connections as shown in [86].

2.3. Cooley-Tukey Algorithm. The Cooley-Tukey FFT algorithm [27] is one of the most important algorithms in the 20th century [21]. It aims to compute the discrete Fourier transform (DFT) of a signal $\{x_n\}_{n=0}^{N-1}$ given by

$$\hat{x}(k) = \sum_{k=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk}, \tag{2.17}$$

in $N \log N$ time. The algorithm leverages the algebraic structure of the N-th complex roots of the unit to recursively split the computation. The simplest version of the algorithm is called the radix-2 decimation-in-time FFT, which computes the DFT of both even-indexed and odd-indexed inputs, which are then merged to produce the final result. In particular, for the first level the DFT is rearranged as

$$\begin{aligned}\hat{x}(k) &= \sum_{m=0}^{N/2-1} x_{2m} e^{-\frac{2\pi i}{N/2} mk} + e^{-\frac{2\pi i}{N} k} \sum_{m=0}^{N/2-1} x_{2m+1} e^{-\frac{2\pi i}{N/2} mk}, \\ &= \hat{x}_e(k) + e^{-\frac{2\pi i}{N} k} \hat{x}_o(k),\end{aligned}$$

where $\hat{x}_e(k)$ and $\hat{x}_o(k)$ stand for the even and odd downsampled DFTs respectively. However, given that we are using decimated DFTs this expression is only valid for $k = 0, ...N/2 - 1$. Thus, in order to obtain the full length DFT, one can use the periodicity of the complex exponential, and we have that

$$\hat{x}(k) = \hat{x}_e(k) + e^{-\frac{2\pi i}{N} k} \hat{x}_o(k), \tag{2.18}$$

$$\hat{x}(k + N/2) = \hat{x}_e(k) - e^{-\frac{2\pi i}{N} k} \hat{x}_o(k). \tag{2.19}$$

2.4. Wide-Band Butterfly Algorithm. For the sake of simplicity we motivate the idea behind this paper, which is the multi-scale decomposition of the butterfly factorization, by using the Cooley-Tukey FFT algorithm. We point out that the same argument can be obtained from a rather involved analysis of the original butterfly algorithm. In particular, one can follow the description of the algorithm in [31] to show that if we build a compressed FIO, as the one in (2.13), at frequency ω using the butterfly algorithm, then most of the computation can be reused to build the same FIO, but at frequency $\omega/2$.

The cornerstone of the approach is to leverage the recursive nature of the FFT algorithm to reuse most of the algorithm pipeline when computing the FFT of decimated signals, or in the case of (2.13) at lower frequencies. We focus our attention on two operations: computing the DFT of a decimated signal using the FFT for a non-decimated signal, computing the same DFT using a decimated algorithm, but

keeping a non-decimated resolution. These two operations will be key when designing our network.

From (2.18) and (2.19) we clearly see that we can compute the DFT of a decimated signal, using the regular FFT algorithm. One only needs to interweave the original signal with zeros, then apply the FFT for the longer signal, and then truncate half of the resulting vector. This means that after a modification of the input we can reuse the algorithmic pipeline from a non-decimated FFT.

Furthermore, if we compute the DFT of a decimated signal, but want to keep the full frequency resolution of the non-decimated one, then (2.18) and (2.19) provides an answer to that: one needs to repeat the result from the decimated signal. This *upscaling* operation will be key when designing the network in Section 3.

These operations follow the same principle behind the wide-band butterfly network. If we want to implement (2.12), we would need to build a network to process the data at each frequency independently. However, using the argument above one can use the recursive decomposition to process the frequencies jointly. In particular, if we want to process data, say at half frequency, i.e., $\omega/2$, then the complementary low-rank conditions states that $d_x d_y \leq 2\omega^{-1}$. If we suppose, in addition, that the evaluation grid remains constant[8] then d_y can be twice as large, thus inducing a different factorization. However, as mentioned above, each factor in H^ℓ factor in the butterfly factorization (see (2.14)) down-samples the local representation in y, while increasing the resolution in x. This means, that after a small modification at the beginning, followed by an upscaling operation similar to the one in (2.18) and (2.19) when the odd signal is zero, one can reuse the rest of the network, which is idea behind merging the networks to treat the different frequencies jointly at the appropriate scale.

3. WideBNet Architecture. We provide a self-contained overview of the network architecture in this section. This material is tailored towards a machine-learning audience with no prior exposure to the butterfly factorization. Indeed, beyond the salient aspects which we summarize below, implementing WideBNet becomes essentially algorithmic since the network structure and connectivity are determined once the dimensions, i.e., grid size, of the data are specified. Our discussion and numerical results consider only two-dimensional scattering. In principle the implementation of our architecture in higher dimensions is straightforward as it is essentially prescribed by the corresponding higher-dimensional butterfly factorization. However, we leave the exploration of WideBNet to three-dimensional inverse scattering, and its attendant complications, to future work.

We separate the discussion into the following. In Section 3.1 we define the sampling and formatting of the input data. Section 3.2 provides the overarching ideas of the architecture and the layers which comprise it. Details about these layers are further elaborated in their respective sections §3.3, §3.4, and §3.5. Lastly, in Section 3.6 we discuss the number of parameters (i.e., trainable weights) present in the network. The pseudo-code for WideBNet is provided in Listing 1 below, whereas the pseudo-codes for the specialized layers V^ℓ, H^ℓ, G^ℓ, and U^L are located in their corresponding subsection[9]. Additionally a depiction of the WideBNet architecture for $L = 4$ levels is shown in Fig. 4

[8] This assumption is a direct consequence of (2.12), where the resolution of the perturbation to be reconstructed is fixed.

[9] We use the notation `LC1D[a,b,c]=LocallyConnected1D(filters=a,kernel_size=b,strides=c)` throughout.

```
def wbnn(Λ^L, ..., Λ^{L/2}):
  y = V^L(Λ^L)
  for ℓ in range(L − 1, L/2 − 1, −1):
    y = H^ℓ(y, V^ℓ(Λ^ℓ))
  y = SwitchResnet(y)
  for ℓ in range(L/2, L, +1):
    y = G^ℓ(y)
  y = U^L(y)
  y = CNN(y)
  return y
```

LISTING 1

Pseudo code for the WIDEBNET *, where each module is explained in detail in Sections 3.4, 3.3, and 3.5.*

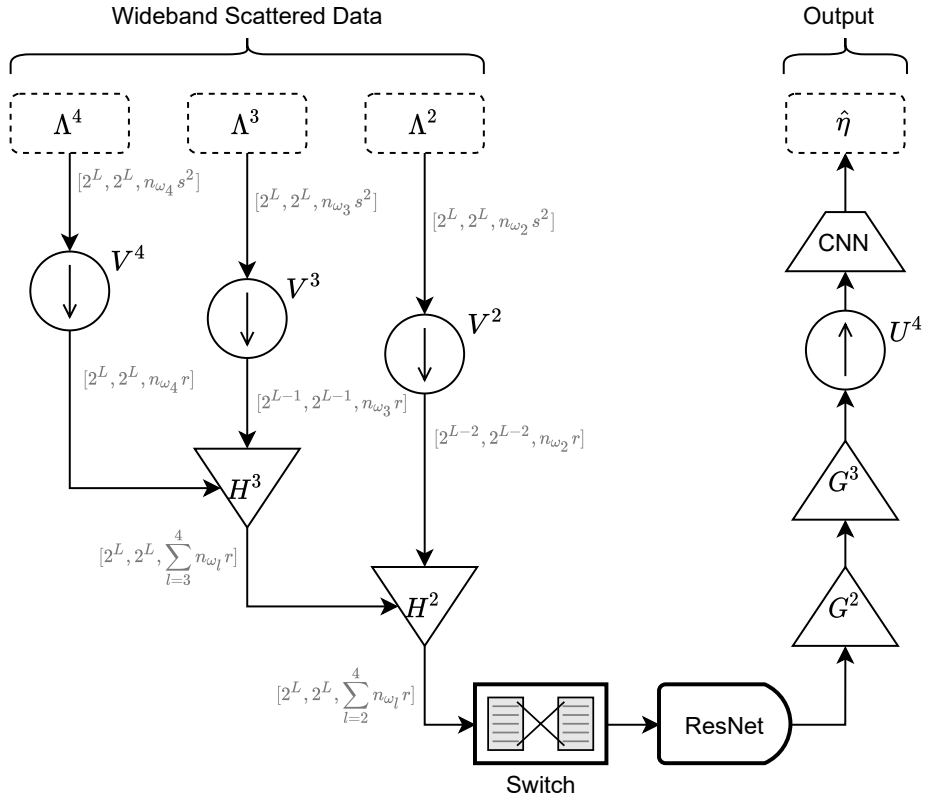

FIG. 4. *Diagram of* WIDEBNET *for data with $L = 4$ levels, leaf size s, and rank r.*

3.1. Input formatting. We assume the scatterers (discretized over an $n_x \times n_z$ grid) and the scattered data (an $n_{\text{src}} \times n_{\text{rcv}}$ matrix for each frequency ω) are represented using complete quad-trees with L levels[10] with leaf size s. In other words,

[10] We require that L is divisible by 2. This is a minor restriction and can be accommodated by e.g. zero padding of the data or by interpolating the data. While the total depth of both quad-trees

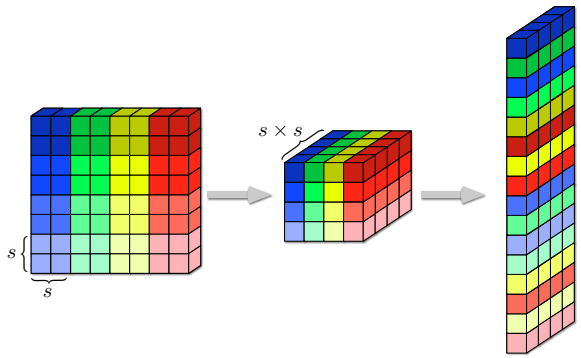

FIG. 5. *Transformation from the image of dimensions* $[2^\ell s,\ 2^\ell s]$ *to the tensorized form of size* $[2^\ell,\ 2^\ell,\ s^2]$*, and then to the flatted tensor using Morton ordering resulting on a tensor of dimensions* $[4^L,\ s^2]$.

we require a discretization into $n = 2^L s$ points for each matrix dimension. The choices of L and s are informed by the inherent wavelengths and sampling frequencies of the inverse problem, and are chosen so that each $s \times s$ voxel of the data matrix Λ^ω are non-oscillatory, i.e., contain at most a few oscillations.

Following the Tensorflow convention of `[height, width, channels]` we reshape these quad-trees into three-tensors of size $[2^L,\ 2^L,\ s^2]$ as shown in Fig. 5. The first two dimensions of the tensor index the geometrical location of the voxels, and the last dimension corresponds to their local vectorial representation. In fact, the data describing the local representation inside each voxel correspond to *channels*. We refer to slices along the height and width dimensions, i.e., the geometrical dimensions, as *patches*. For example, a 1×1 patch of data describes slices of the three-tensor with dimension $[1,\ 1,\ s^2]$. As we discuss shortly, at the finest spatial resolution WIDEBNET operates on 1×1 patches, and at the coarsest spatial resolution it operates on $2^{L/2} \times 2^{L/2}$ patches. It is convenient to introduce levels $\ell \in [L/2,\ L]$ to index the resolution, or equivalently, the size of the contiguous $2^{L-\ell}s \times 2^{L-\ell}s$ sub-matrices in the data matrix that will be processed.

For the purpose of describing our network using linear algebraic operations it is convenient to characterize these three-tensors as equivalently reshaped two-tensors of size $[4^L,\ s^2]$. This flattening proceeds according to a natural ordering of quad-trees known as "Morton-ordering" or "Z-ordering", which is depicted in Fig. 5. We refer to [60] for more details.

As we discussed in the introduction it is beneficial for the stability of the inverse problem for the input data $\Lambda^\omega \in \mathbb{C}^{n_{\text{src}} \times n_{\text{rcv}}}$ to be collected from a wideband of frequencies $\omega \in \Omega = [\omega_{\text{low}},\ \omega_{\text{high}}]$. The bandwidth ω_{low} and ω_{high} is determined from the experimental configuration. For our data assimilation strategy we index this bandwidth with a dyadic partition containing $L/2 + 1$ intervals: for $L/2 \le \ell \le L$ we label the intervals $\Omega^\ell = (\omega_{\text{low}} + 2^{L-\ell-1}\Delta\omega,\ \omega_{\text{low}} + 2^{L-\ell}\Delta\omega]$ where $\Delta\omega = 2^{-L}(\omega_{\text{high}} - \omega_{\text{low}})$. We assume that within each interval Ω^ℓ we probe the medium with n_ω^ℓ frequencies, not necessarily equi-spaced, and with slight abuse of notation denote the resulting dataset as $\Lambda^\ell \in \mathbb{C}^{n_{\text{src}} \times n_{\text{rcv}} \times n_\omega^\ell}$. Following the quad-tree structure we reshape each data tensor Λ^ℓ into a three-tensor of size $[2^\ell,\ 2^\ell,\ n_\omega^\ell s^2]$ by concatenating all the

must be the same, it is not necessary for them to have the same leaf size. However, for ease of presentation, our discussion focuses exclusively on this case.

multi-frequency data collected from bandwidth Ω^ℓ along the channel dimension. The input to WideBNet thus consists of the collection of scattering data $\{\Lambda^\ell\}_{L/2\leq\ell\leq L}$.

3.2. Architecture Overview. We aim to incorporate the physics of wave propagation into the design of our network by translating analytic properties of the discrete imaging condition (2.12) into neural modules. Since the imaging condition is derived by linearization of the partial differential operators in the wave equation, this process should, at minimum, ensure that our network is able to capture the physics of single wave scattering. To that end, for a set of given frequencies $\{\omega_\ell\}_{\ell=L/2}^{L}$ we seek an architecture that can emulate the functionality of the imaging algorithm

$$\{\Lambda^\ell\} \mapsto \sum_{\ell=L/2}^{L} \alpha(\omega_\ell) \left((F^{\omega_\ell})^* F^{\omega_\ell} + \epsilon I\right)^{-1} (F^{\omega_\ell})^* \Lambda^\ell. \tag{3.1}$$

We emphasize, however, that we are ultimately interested in applications of WideBNet to data beyond the Born single scattering regime associated with the imaging condition.

We leverage the following analytic properties of the imaging condition. First, as originally elucidated in [51], we recognize that for each frequency ω the regularized normal operator $\left((F^\omega)^* F^\omega + \epsilon I\right)^{-1}$ corresponds to a translation invariant operator. Second, we also recognize that the operator F^ω describes a generalized Fourier operator which is amenable to a butterfly factorization. In other words, after suitable discretization the operator F^ω admits a matrix decomposition viz.,

$$F^\omega = U^L G^{L-1} \dots G^{L/2} S^{L/2} H^{L/2} \dots H^{L-1} V^L. \tag{3.2}$$

In the traditional linear setting of the discrete imaging condition with Morton-flattened data we have $U^L \in \mathbb{C}^{4^L s^2 \times 4^L r}$, $V^L \in \mathbb{C}^{4^L r \times 4^L s^2}$, and all other remaining matrix factors of dimension $4^L r \times 4^L r$. Most importantly, each matrix factor in the butterfly decomposition has a sparsity pattern that is informed by analytic considerations of the wave kernel. These sparsity patterns in the matrix become equivalent to a block diagonal operator after specific permutation of either the columns or the rows.

WideBNet utilizes these two insights to replace the functionality of $\left((F^\omega)^* F^\omega + \epsilon I\right)^{-1}$ and F^ω by analogous neural modules. We translate the butterfly decomposition for the generalized Fourier operator F^ω by replacing each matrix factor (e.g. $U^L, V^L, \dots$) by neural network layers[11]. In this setting the permutation and sparsity structure of the butterfly matrix factors inform the inter-layer network connectivity (i.e., the network topology), and the matrix entries themselves become trainable weights in the network. The information processed by these layers are ultimately sent data into a CNN module, which are well adapted to capture translation invariant operators, and thus mimics the effects of the regularized pseudo-inverse in sharpening the estimate coming from the imaging condition.

If only monochromatic data are considered, e.g. using solely scattering data $\Lambda^L \in \mathbb{C}^{n\times n\times n_\omega^L}$ obtained by probing the medium at only $n_\omega^L = 1$ frequency, then the network just described is equivalent to other butterfly-based networks BNet [58] or SwitchNet [51]. However, rather than replacing each ω-dependent operator in the imaging condition (3.1) with individual butterfly-based networks, WideBNet instead

[11] For ease of comparison we retain the transpose $\cdot^*$ in the naming convention of our network but note that transposition is no longer meaningfully defined in our new non-linear setting.

aims to more efficiently assimilates multi-frequency data with the following modifications:

(i) We exploit the connection between spatial resolution and frequency in wave-scattering problems by processing data only at their relevant length scales. We note that each H^ℓ layer, analogous to their butterfly factorization namesakes, processes data over voxel patches of size $2^{L-\ell} \times 2^{L-\ell}$, i.e., the effective length scales at this layer are of order $2^{L-\ell}$. As a result, the dispersion relation in wave-scattering suggests that data from bandwidth Ω^ℓ are most informative at this length-scale[12], and thus we feed in data accordingly. This strategy of dyadically partitioning the bandwidth to localize spatial information is also employed by the Cooley-Tukey FFT algorithm to achieve quasi-linear time complexity [27]; in our setting this strategy affords us significant reductions in the number of trainable weights in the network.

(ii) In addition to the switch permutation layer, we also introduce non-linearities into the network using residual network which we call the SWITCH-RESNET layer. Information from the entire bandwidth of data is thus processed at this layer. These non-linearities, in theory, extend the functionality of WIDEBNET beyond the limitation the discretized imaging condition; namely, the implicit assumption of Born single scattering. Furthermore, non-linear combinations of wideband data are known to be a strict requirement of super-resolution imaging [32], and therefore potentially enabling WIDEBNET image estimates to achieve resolutions below the Nyquist limit.

In the following sections we elaborate on the specific details of each specialized butterfly-network layer in Alg. 1.

3.3. U^L and V^ℓ layers. In the traditional numerical analysis setting the butterfly matrix factor V^ℓ in (3.2) represents a block diagonal matrix with block size $r \times s^2$. This operator takes input data (viewed as a complete quad-tree) and compresses leaf nodes at level L, each with $s \times s$ degrees of freedom, into 1×1 patches with $\sqrt{r} \times \sqrt{r}$ degrees of freedom; this process is depicted in Fig. 6. Similarly, the U^L matrix factor in (3.2) is also block diagonal but instead with block sizes of $s^2 \times r$. This operator thus "samples" the local representation of dimension r back to its nominal dimensions of $s \times x$. In both instances the compression/decompression is essentially lossless provided the number of levels L is properly adapted to the probe frequency ω. We emphasize again that this follows as a consequence of the *dispersion relation* in wave-scattering: provided these parameters are chosen correctly, then over $s \times s$ length scales the data are non-oscillatory (i.e. sub-wavelength) and therefore admits a low-rank representation with rank r.

WIDEBNET also exploits this relation between spatial resolution and frequency. However, a key point of departure from the butterfly factorization is that here the input data are wideband and thus contains multiple length scales (wavelengths). This motivates the introduction of auxiliary layers V^ℓ for $L/2 \leq \ell \leq L$ whose inputs are assumed to be sampled from bandwidth Ω^ℓ. Each V^ℓ layer compresses the input data *at level* ℓ such that nodes with $2^{L-\ell}s \times 2^{L-\ell}s$ degrees of freedom are mapped into 1×1 patches with $\sqrt{r} \times \sqrt{r}$ degrees of freedom; this also has the interpretation of spatial downsampling. Note that the dyadic scaling in the definition of Ω^ℓ is critical in maintaining the balance between spatial resolution and frequency.

[12]The $\{G^\ell\}$ layers have a similar multi-resolution property. This suggests that data from bandwidth Ω^ℓ should also be fed into G^ℓ similar to the U-Net [78] architecture; however, numerical results demonstrate that this additional complexity is unnecessary.

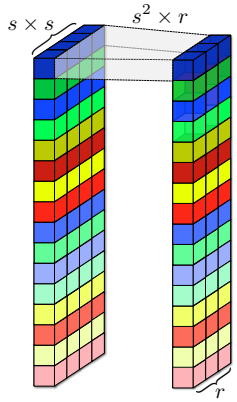

FIG. 6. *Sketch of the compression carried in the V^L layer, from the points contained in a leaf of size $s \times s$ (see Fig. 5) to a local representation of rank r. The grey polygon represent the connections between the two layers.*

When the input data Λ^ℓ from bandwidth Ω^ℓ are represented as a three-tensor of dimension $[2^\ell,\ 2^\ell,\ n_\omega^\ell]$, each V^ℓ layer can be implemented as a `LocallyConnected2D` layer in Tensorflow with rn_ω^ℓ channels and both the kernel size and stride as $2^{L-\ell} \times 2^{L-\ell}$. The U^L layer can also be implemented as `LocallyConnected2D` layer with rank s^2 and 1×1 kernel size and stride; the input to this layer is assumed to be of dimension $[2^\ell,\ 2^\ell,\ c]$ with c input channels. For completeness, we provide the implementation of these layers in Algs 2 and 3 for input data that is Morton-flattened. Furthermore, note the pseudo-code also details the processing when input data contain both real and imaginary components.

```
def V^ℓ(X):
  # input:  [?, 2, 4^ℓ, 4^{L−ℓ} s^2 n_{ω_ℓ}]
  # output:  [?, 2, 4^L, r n_{ω_ℓ}]
  y_re, y_im = X[:,0,:,:], X[:,1,:,:]

  x_re = LC1D[r n_{ω_ℓ},1,1](y_re) + LC1D[r n_{ω_ℓ},1,1](y_im)
  x_im = LC1D[r n_{ω_ℓ},1,1](y_re) + LC1D[r n_{ω_ℓ},1,1](y_im)

  return tf.stack([x_re, x_im], axis=1)
```

LISTING 2
Pseudo code for the V^ℓ module for Morton-flattened data.

```
def U^L(X):
  # input:  [?, 2, 4^L, ·]
  # output:  [?, 2, 4^L, s^2]
  y_re, y_im = X[:,0,:,:], X[:,1,:,:]

  x_re = LC1D[s^2,1,1](y_re) + LC1D[s^2,1,1](y_im)
  x_im = LC1D[s^2,1,1](y_re) + LC1D[s^2,1,1](y_im)

  return tf.stack([x_re, x_im], axis=1)
```

LISTING 3
Pseudo code for the U^L module for Morton-flattened data.

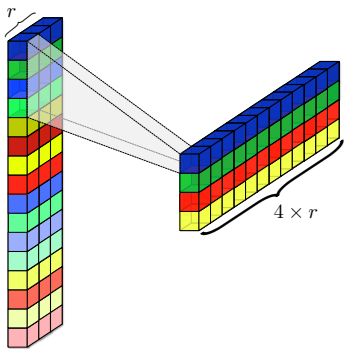

FIG. 7. *Sketch of the application of the H^ℓ layer. The layer decimates by a factor of four the number of neurons in the spatial dimension, while increasing four times the number of channels. From Fig. 5 we can observe that the decimation is equivalent to decimate by a factor two in each of the first two dimensions, which follows from the Z-ordering.*

Remark: A major application of the butterfly factorization is for applying FIOs in linear-time complexity; the rank r then depends on the error tolerance but generally requires that $r \ll s^2$. This represents a significant philosophical difference in how in r is determined in our machine learning setting – it does not matter if $r \geq s^2$ so as long as the learned model achieves its intended task. Nevertheless, our numerical results in §4.1.4 demonstrate that *(i)* it suffices to choose $r \ll s^2$ and, moreover, that *(ii)* generalization is largely insensitive to the choice of r.

3.4. H^ℓ and G^ℓ layers. The H^ℓ and G^ℓ factors in (3.2) continue the theme of multi-scale processing. When viewed as matrices, both H^ℓ and G^ℓ are block diagonal with block size $4^{L-\ell}r \times 4^{L-\ell}r$. Equivalently, when the input is formatted as a complete quad-tree, this implies both are *local operators* which process the nodes on the tree at length scale l to map each $2^{L-\ell}s \times 2^{L-\ell}s$ patches. Within each block there is further structure to the operators, as Figure 7 demonstrates. For each H^ℓ each sub-block has the interpretation of *aggregating* information, whereas each G^ℓ achieves the dual task of *spreading* information. We stress, however, that the action of this is entirely local within each patch. In either case, the key observation is that by permuting each node following a set pattern each operator becomes block-diagonal with block size 4^ℓ, for all $L/2 \leq \ell \leq L$. The specific permutation pattern π_ℓ enabling this matrix partitioning is discussed in Appendix A.

In our WIDEBNET adaptation, each G^ℓ layer directly mimics the behaviour of their counterparts and can be implemented using the LOCALLYCONNECTED2D layer with 4×4 kernel sizes and stride 4. The number of channels is chosen to be $\sum_{i=\ell+1}^{L} rn_{\omega_i}$ for symmetry.

However, note that our H^ℓ layers require modification on account of our data assimilation strategy to inject information at their correct length scales. As such, these layers process two inputs: one the output of the V^ℓ layer of dimension $[2^{L-\ell}, 2^{L-\ell}, rn_{\omega_\ell}]$, the other the output from the previous layer of dimension $[2^\ell, 2^\ell, c]$ for some channel size c. To process the dimensions of both we first upscale each patch with redundant information to convert the data into $[2^\ell, 2^\ell, rn_{\omega_\ell}]$. This is then concatenated with the other input to form a tensor of size $[2^\ell, 2^\ell, c + rn_{\omega_\ell}]$. Note that the ordering of the concatenation along the channel dimension does not matter so as long as it is performed consistently.

Alg 4 provides a pseudo-code implementation of the H^ℓ layer when using Morton-flattened inputs.

```
def H^ℓ(X, W):
  # input X:  [?, 2, 4^ℓ, ·] from V^ℓ(Λ^ℓ)
  # input W:  [?, 2, 4^L, ·] from previous layer
  # output:  [?, 2, 4^L, r Σ_{i=ℓ}^{L} n_{ω_i}]

  X̃ = UpSampling2D(X)

  X = tf.stack([X, X̃], axis=-1)

  y_re, y_im = X[:,0,:,:], X[:,1,:,:]

  # permute patches according to π_ℓ before
  # applying block diagonal transformations
  y_re = y_re[:,π_ℓ,:]
  y_im = y_im[:,π_ℓ,:]

  x_re = LC1D[4r Σ_{i=ℓ}^{L} n_{ω_i},4,4](y_re) + LC1D[4r Σ_{i=ℓ}^{L} n_{ω_i},4,4](y_im)
  x_im = LC1D[4r Σ_{i=ℓ}^{L} n_{ω_i},4,4](y_re) + LC1D[4r Σ_{i=ℓ}^{L} n_{ω_i},4,4](y_im)

  return tf.stack([x_re, x_im], axis=1)
```

LISTING 4
Pseudo code for the H^ℓ module for Morton-flattened data.

3.5. Switch-Resnet layer. We retain the permutation pattern of the switch layer as this is responsible for capturing the inherent non-locality of wave scattering (e.g. a point scatterer generates a diffraction pattern that is measured by all receivers in our geometry). We illustrate this pattern in Fig. 8, and the specific description of the permutation indexing π_{switch} can be found in Appendix A.

The input to this level serves as a condensed representation of the measured data. It is at this level that we non-linearly process the multi-frequency dataset; we speculate that this also essential in facilitating the model to produce super-resolved images. We achieve this using a residual network to refine each channel locally following each resnet unit. The pseudocode is provided in Alg. 5.

```
def SwitchResnet(X):
  # apply switch local-to-global permutation
  y = X[:,:,π_switch,:]

  # non-linear synthesis
  # apply switch permutation
  for k in range(N_resnet):
    y = LC1D[r,1,1](ReLu(LC1D[r,1,1](y)) + y
    if k < N_resnet:
      y = ReLu(y)
```

LISTING 5
Pseudo code for the SWITCH-RESNET *module for Morton-flattened data.*

3.6. WideBNet Parameter Count. An estimate of how the number of parameters (i.e. trainable weights or degrees of freedom (d.o.f.)) scales is

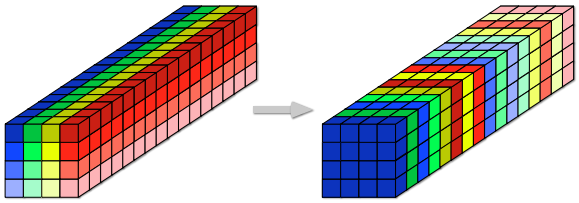

FIG. 8. *Visualization of the transformation by the switch permutation layer for an example with $r = 1$ and $L = 4$. For this configuration the input and output are both tensors of dimension $[2^{L/2}, 2^{L/2}, r4^{L/2}] = [4, 4, 16]$. The local information contained at each geometric position (equivalently, the channel information) is distributed globally according to the switch permutation pattern π_{switch}.*

$$\text{d.o.f.}(\text{WIDEBNET}) \approx 4^{\ell} r^2 \left(\sum_{l=L/2}^{\ell} n_{\omega_\ell}^2 + \sum_{l=L/2}^{\ell} \left(\sum_{i=l}^{\ell} n_{\omega_i} \right)^2 \right).$$

When only a single frequency is sampled in each sub-band, i.e. $n_{\omega_\ell} = 1$ for all l, then this total becomes $\mathcal{O}\left(N(\log N + \log^3 N)\right)$. Note this is essentially linear in the total degrees of freedom in the data (N) up to poly-logarithmic factors. Furthermore, note if naïvely L separate single channel WIDEBNET networks were used to compute (2.12) this would correspond to complexity $\mathcal{O}\left(N \log N^2\right)$; the multi-frequency assimilation only exceeds this with mild oversampling by a logarithmic factor.

Lastly, we note the effect of the partitioning of the frequencies. If all the frequencies were ingested at length scale L then the scaling becomes $\mathcal{O}\left(N(\log N^2 + \log N^3)\right)$. While to leading order this presents the same asymptotic scaling, in terms of practical considerations this presents as substantial increase in the number of trainable parameters.

4. Numerical Results. Synthetic data were generated using numerical finite differencing for (2.3) over the computational domain $[-0.5, 0.5]^{\otimes 2}$. The domain was discretized with an equispaced mesh of $n_x = 80$ by $n_z = 80$ points which corresponds to a quad-tree partitioning into $L = 4$ levels with leaf size $s = 5$. Training data were generated using a second-order finite difference scheme while *testing* data were computed with fourth-order finite differences. The use of higher quality simulations for testing serves to validate that WIDEBNET predictions do not depend on computational artifacts such as e.g. numerical dispersion. The radiating boundary conditions for Eq 2.3 were implemented using perfectly matched layers (PML) with a quadratic profile with intensity 80 [8]. The width of the PML was chosen to span at least one wavelength at the lowest frequency.

Unless specified otherwise the dataset consisted of $n_f = 3$ source frequencies at 2.5, 5, and 10 Hz. In a homogeneous background with velocity $c_0 = 1$ this corresponds to 8 points-per-wavelength (PPW) at the highest frequency. Receivers were located at equi-angular intervals around a circle of radius $r = 0.5$ with the recorded data computed by linearly interpolating the scattered field. We used $N_{\text{rcv}} = 80$ receivers and sources for all experiments. For a homogeneous background the direct wave is given analytically (see (2.3)). In these instances the directions of arrival $\mathbf{s} \in \mathbb{S}^1$ were aligned with the receiver geometry, i.e. incident from 80 equiangular directions. However, for inhomogeneous media the direct waves had to be computed numerically. This

was achieved by using numerical Dirac deltas as source functions. These sources were localized on a circle of radius $r = 1$ at 80 equiangular intervals and the computational domain was extended to $[-1, 1]^{\otimes 2}$ using the same grid spacing Δx and Δz as before. The resulting scattered field was computed by differencing the solutions to (2.3) with and without scatters. The acquisition geometry was fixed for all frequencies.

Scatterers were selected from a dictionary of simple, convex, geometric objects such as squares, triangles, and Gaussian bumps. The characteristic lengths of the square and triangular scatterers were measured with respect to their base, rather than the diameter of the smallest enclosing ball, whereas the characteristic length of the Gaussian was taken to be its standard deviation. In each data point the number of scatterers was determined by uniformly sampling from $\{2, 3, 4\}$ objects, and their locations were uniformly distributed inside a circle of radius $r = 0.35$. No restrictions were enforced against overlapping scatterers. In all experiments the amplitude of each scatterer was fixed to 0.2; we leave to future work how the training data can be augmented to account for variations in amplitudes.

WideBNet was implemented in Tensorflow [1] and trained with the pixel-wise sample loss function

$$\sum_{x=\text{pixel in image}} \left\| (K_{\text{high}} * \eta)(x) - \text{WideBNet}\ [\Lambda^L, \ldots, \Lambda^{L/2}](x) \right\|_2^2, \tag{4.1}$$

where η denotes the sample realization of the scatterer wavefield and $\{\Lambda^\ell_{s,r}\}_{L/2 \le l \le L}$ the partitioned multi-frequency data. This objective function was chosen to promote the recovery of an image that is *smoother* than the true numerical solution by a factor of a two-dimensional convolution with high-pass filter K_{high}. Critically we still remain in the super-resolution regime when the support of filter K_{high} is significant smaller than the Nyquist limit of $\lambda_{min}/2$[13] as the smoothed image still contains sub-wavelength features. This strategy was inspired by the work of [17] who relied on this insight for theoretical proofs on recoverability limits in super-resolution. In our experiments we selected K_{high} to be a Gaussian kernel with characteristic width of 0.75 grid points (compare this the diffraction limit in our bandwidth of 4 pixels). This smoothing was observed to be integral in promoting stable training dynamics. We also report the image-wise relative error

$$\frac{\left\| K_{\text{high}} * \eta - \text{WideBNet}\ [\Lambda^L, \ldots, \Lambda^{L/2}] \right\|_2^2}{\|K_{\text{high}} * \eta\|_2^2}. \tag{4.2}$$

Note that we do not normalize the norms in either (4.1) or (4.2) by the grid lengths Δx and Δz.

The dataset was split into 21000 training points and 4000 testing points[14], respectively, with batch size 32. Note, in comparison, an instance of WideBNet with $N_{\text{CNN}} = 3$ convolutional layers and $N_{\text{RNN}} = 3$ residual layers contains 200000 trainable parameters meaning our models are still in the massively over-parameterized regime. Unless specified otherwise the testing set follows the same distribution (e.g. scatterer types) as the training set. The initial learning rate (i.e. step size) was universally set to 5e-3 across all experiments. The learning schedule was set according to Tensorflow's [1] implementation of `ExponentialDecay` with a decay rate of 0.95

[13] The ratio of these two quantities is the so-called *super-resolution factor*.

[14] a single "data point" has dimension $n_x \times n_z \times n_f$

after every 2000 plateaus steps with stair-casing. We chose the Adam optimizer [52] and terminated training after 150 epochs. No special initialization strategy was required and the network weights were randomly initialized with `glorot_uniform` – we did not observe the training instabilities with random initialization that were thoroughly documented in [86] for general butterfly networks. All computations were done with `float32` half-precision. Note that no effort was taken to optimize these hyper-parameters using an external validation set.

4.1. Homogeneous Background. In this section we present numerical results for WideBNet models trained with scattered data that propagated through a known homogeneous background medium of wavespeed $c_0 = 1$. Each row of Figure 12 depicts WideBNet predictions on testing data across a variety of scatterer configurations. Except for Figure 12c the data were sampled from the bandwidth of 2.5, 5 and 10 Hz which implies a limiting wavelength of 8 points per wavelength (PPW). Figures 12a and 12b involve a multi-scale dictionary of scatterers with characteristic lengths ranging from 3, 5, and 10 pixels; these correspond to the sub-wavelength, wavelength, and super-wavelength regimes, respectively. We observe that WideBNet correctly localizes each scatterer in addition to resolving sub-wavelength features such as e.g. the corners of the triangles. Figure 12d similarly depicts a heterogeneous dictionary but with rotated triangles of fixed side-length 5 pixels. In Figure 12c the same experiment was repeated but with a bandwidth that was shifted to 1.25, 2.5 and 5 Hz so that the limiting wavelength increases to 16 PPW; in this regime all scatterers are sub-wavelength. Nevertheless, WideBNet still produces images that are qualitatively comparable to the higher bandwidth experiments. This suggests that our algorithm has a high super resolution factor. For completeness, we include results in Figure 12e for point scatterers that were originally proposed for super-resolution by Donoho [32].

Table 2 summarizes the training and testing loss for various scatterer configurations. Each row corresponds to a separate experiment with triangular ($\triangle$), square ($\square$), or Gaussian ($\bigcirc$) scatterers. The numbers in the parentheses correspond to the characteristic length, in pixels, with multiple numbers indicating a multi-scale dataset.

Several trends can be observed from this table. In all configurations there is no evidence of over-fitting; indeed, the generalization gap, defined to be the difference between the testing and training errors, is on average less than an order of magnitude. Furthermore, both qualitatively and quantitatively there is no significant difference between datasets with a fixed characteristic length versus the multi-scale datasets. This demonstrates robustness to the choice of the scatterer dictionary. However, we observe that Gaussian scatterers outperform other shapes across all metrics, perhaps owing to their smoothness. Overall, the pixel-wise error in testing tends to decrease with decreasing length scale; we conjecture the exact scaling may depend on the perimeter to area ratio of the polygons.

4.1.1. Effect of Switch Layer. In Section 3 we emphasized the importance of the switch permutation pattern in representing the local-to-global physics of wave scattering. Figure 14 corroborates this claim by comparing the predictive ability of WideBNet models trained with and without the inclusion of the switch permutation layer. Both models contain the same number of trainable weights, and all other configurations were held equal.

Figure 14 demonstrates that the predictions *without* the permutation layer are of noticeably poorer quality. However, the switch-less configuration manages to localize scatterers and even reproduces sub-wavelength features to an extent, particularly when the scatterers are well separated as in Figure 14(b). However, Figures 14(c) and

(d) exposes the deficiencies of this model in the presence of overlapping scatterers, i.e., in the super-resolution regime where scatterers are separated by sub-wavelength distances. We observe that these complications appear to be remedied by the inclusion of the switch permutation layer.

Although the switchless configuration manages to produce reasonable images, we conjecture that this is because the model is "reasonably deep" at this length scale. We suspect the predictive abilities will quickly deteriorate as $L \to \infty$ since the depth of the network only scales linearly as $\Theta(L)$.

4.1.2. Out-of-Distribution Generalization. We consider the performance of WideBNet on scatterers that are *out-of-distribution* and distinct from the *within-distribution* scatterers of the training set. The result of this experiment is shown in Fig. 9 for a WideBNet model that is trained with randomly located squares of sidelengths 3, 5, and 10 pixels. An example datapoint from this training class is presented in the left-most column. This trained network is applied to four different scattering configurations: a non-convex shaped 'blob', shown in the top row of the second column; Gaussians with characteristic length of 2 pixels, shown in the third column; triangles with sidelengths of 3 and 10 pixels, shown in the fourth column; and a Shepp-Logan phantom, shown in the last column. All colour scales are normalized with respect to the first row, which corresponds to the exact solution.

The second row of the figure depicts the output of WideBNet from noiseless, wideband, recordings of the respective scattered wavefields. We observe that our network generalizes to two distinct, forward scattering, regimes. First, WideBNet is able to localize both the Gaussian and the triangular scatterers, in addition to resolving their shapes and sub-wavelength features. This suggests that our network learns an inverse scattering map applicable to general configurations that are dominated by Born single scattering, with seemingly no limitations on resolution. Second, we note WideBNet is also capable of generalizing to data involving strong multiple scattering, as indicated by its performance on the multiple wavelength spanning and non-convex 'blob'. As Fig. 9 demonstrates, WideBNet infills the shape, although with noticeable errors along the support of the scatter. Since during training it is only provided with data with at most three square scatterers, this infilling property suggests that our model captures some generalized properties of inverse wave scattering. We note, however, there are limits to its extrapolative ability, as suggested by the results involving the Shepp-Logan phantom. Characterizing *a priori* which configurations are amenable to extrapolation remains an open problem.

Remarkably, other examples of neural networks extrapolating beyond the scattering configurations of their training sets have been reported in the literature. In [71] the authors apply an LSTM network, designed to explicitly incorporate the Lippmann-Schwinger kernel, to learn the physical model which generates the wavefield from scatterers. They report the ability of their network to simulate wavefields from scattering shapes unseen in training. In a similar vein, [55] consider the inverse scattering problem, with a network architecture called FIONet which also leverages principles from Fourier integral operators, and successfully image out-of-distribution scatterers. We leave an investigation into this commonly observed extrapolation phenomena to future work.

4.1.3. Partitioning of frequencies. Table 1 reports on the difference between two competing frequency partitioning strategies: "AllFreq", in which the data from the entire bandwidth are fed into WideBNet at level L, versus "MultiFreq" wherein the data are only processed at the appropriate length scale ℓ. Qualitatively both

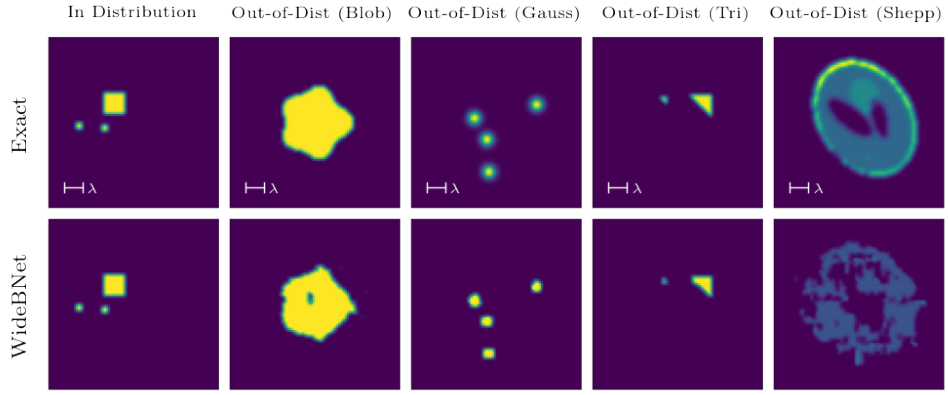

FIG. 9. *Out of distribution performance of a* WIDEBNET *model trained on a Square 3/5/10 dataset (example datapoint shown in left-most column). The second to fifth columns depict examples of out-of-distribution datapoints. All colour scales are normalized with respect to the exact solutions shown in the first row.*

TABLE 1
Effect of frequency partition

		Pixel-wise Squared Loss		Image-wise Relative Loss	
	DOF	Train	Test	Train	Test
AllFreq	2746368	2.92E-06	4.81E-06	1.26E-05	1.72E-05
MultiFreq	1913856	4.06E-06	6.40E-06	1.72E-05	2.27E-05

strategies produce comparable images that are sharp and resolve the sub-wavelength features. In fact, quantitatively the "AllFreq" strategy produces marginally lower losses (though within the same order of magnitude). However, as noted in Table 1 that the degrees of freedom of "AllFreq" far exceed that of "MultiFreq"; although both strategies have the same asymptotic storage complexity of $\mathcal{O}(N \log^3(N))$ (see Section 3.6), practically speaking the constant differs by a substantial amount in favour of "MultiFreq".

We report that we were unable to successfully train a model by mimicking (2.12) directly, i.e. training single channel WIDEBNET models for each frequency independently, then merging their predictions via a CNN module. This is perhaps unsurprising since it is known that super-resolution algorithms require non-linear synthesis of multi-frequency data to succeed. Whereas in both "MultiFreq" and "AllFreq" this is achieved by the switch-resnet module, in this naive strategy the synthesis is performed only at the end by the CNN layers. In comparison to the optimal storage complexity of $\mathcal{O}(N \log^2 N)$ in this naïve strategy, note that mildly overparametrizing by a small logarithmic factor provides significant training stability to the inverse problem.

4.1.4. Training Curves & Hyper-Parameter Sensitivity.

Training Curves. Figure 11(a) reports the training errors for models trained on datasets containing 5000, 10000, 15000, and 21000 datapoints. The trained models were evaluated on a *fixed* testing set of 3000 points (i.e. the same testing set is to

compare all experiments). All remaining hyperparameters such as the learning rate and number of epochs were held the same as discussed in the beginning of Section 4. Note in all cases we remain in the over-parametrized regime since the number of datapoints is far fewer than the number of degree of freedom. Nevertheless, with only a few samples WIDEBNET stably achieves a pixel-wise loss on the order of 10^{-5}.

We observe in Figure 11 that both training and testing errors decrease with increasing training points, as expected. However, these training/testing curves quickly saturate and the differences fall less than an order of magnitude. Furthermore, the empirical generalization gap, taken to be the difference between the testing curve (dashed lines) and the training curve (solid line) remains within the same order of magnitude as the number of points is increased. These points demonstrate that our model *(i)* generalizes with relatively scant training points, and *(ii)* saturates its model capacity quickly, which is an indication that the architecture is well adapted to the task.

Sensitivity to the rank r. While the data essentially specify the architecture through requirements on the level L and leaf size s, it remains up to the user to select the rank r. We reiterate that this choice serves as a significant departure from the numerical analysis perspective of the Butterfly factorization; whereas in the original context it is essential to have the scaling $r \ll s^2$ for the purpose of fast matrix-vector multiplication[15], in the current machine learning context there is no restriction against choosing $r \geq s^2$. Nevertheless, as Figure 11b demonstrates, a large over-parameterization with respect to r is unnecessary. Indeed, while the training metrics monotonically decrease as the model capacity increases with rank, we observe that testing errors remain relatively saturated. This suggests that performance of WIDEBNET is largely insensitive to the rank and the network topology plays a more significant role.

Moreover, these results indicate that allowing for a non-uniform rank for each patch may not yield be a fruitful exercise. Or, conversely, if the intent is to compress the model further to e.g. fit on mobile devices [9], this also suggests that tenable strategy may be to prune a trained model by adaptive patch-wise rank reduction. We leave this to future work.

Effect of CNN and ResNet Layers. Beyond the selection of the rank r the only remaining hyper-parameters that determine the WIDEBNET architecture are the number of CNN layers N_{CNN} and the number of residual layers N_{RNN} in the switch module. Figure 15 reports on the sensitivity of the WIDEBNET model to these parameters. Evidently from Figure 15b we conclude that the predictive performance is unaffected by the number of post-processing CNN layers. A similar conclusion can be drawn about the number of residual layers from Figure 15a; note the fluctuations in the training and testing curves are negligible in magnitude.

4.2. Heterogeneous Background. In this section we present numerical results with scattering data from a known *inhomogeneous* background medium. The variations in the background wavespeed introduce significant complications to the inverse problem. For instance, homogeneous backgrounds afford symmetries such as rotational equivariance which can be exploited for efficient network design, see e.g. [37]; in an inhomogeneous background this assumption is no longer valid. The physics of wave propagation through inhomogeneous media also complicates the signal pro-

[15]Typically the rank is determined by computations of SVDs so that ϵ is close to be machine zero. Analytical relations between ϵ and r are kernel dependent and is known explicitly only in few cases.

TABLE 2

Training and testing errors for various datasets. Each experiment consisted of 21000 training points and WIDEBNET *was evaluated against an independent testing dataset with 3000 points. The data were generated using a homogeneous background wavefield $c_0 = 1$ and data were sampled at 2.5, 5, and 10 Hz (i.e. the effective wavelength was 8 PPW). Each row denotes a separate experiment with the scatterers consisting of triangles (△), squares (□), and Gaussians (◯). The numbers in the parentheses indicate the characteristic lengths of each scatterer; multiple numbers indicate a heterogeneous dataset of scatterer sizes, while (rot,5) indicates a dataset with rotated scatterers. In general, we observe that* WIDEBNET *does not over-fit the data. Surprisingly, on average the testing pixel-wise error decreases with decreasing length-scale.*

	Pixel-wise Squared Loss		Image-wise Relative Loss	
Scatterer	Train	Test	Train	Test
△ (3,5,10)	4.06E-06	6.40E-06	5.38E-04	7.12E-04
□ (3,5,10)	7.12E-04	1.13E-05	4.63E-04	6.24E-04
◯ (3,5,10)	1.24E-06	2.01E-06	1.89E-05	2.71E-05
△ (rot,5)	3.03E-06	4.09E-06	5.52E-04	7.32E-04
△ (10)	2.47E-06	2.51E-05	9.26E-05	8.17E-04
△ (5)	1.14E-06	7.19E-06	2.11E-04	1.24E-03
△ (3)	4.35E-06	4.23E-06	2.62E-03	2.62E-03
□ (10)	2.63E-06	7.92E-05	4.90E-05	1.24E-03
□ (5)	1.24E-06	2.09E-05	1.13E-04	1.75E-03
□ (3)	1.19E-05	1.19E-05	3.77E-03	3.80E-03
◯ (3)	9.89E-08	2.61E-06	5.97E-06	1.30E-04
◯ (2)	3.19E-07	4.84E-07	4.35E-05	6.28E-05
◯ (1)	5.86E-07	7.52E-07	4.71E-04	5.87E-04

cessing problem as it gives rise to multi-pathing as well as multiple arrivals due to interior scattering. While the architecture and data formatting remain unchanged, the complexity of the inverse problem for localizing scatterers, let alone super-resolution, increases in this setting.

We tested the algorithm for two heterogeneous backgrounds: (i) a smooth linearly increasing background medium with wavespeed $c = 0.5$ at the top and $c = 1.5$ at the bottom, and (ii) layered background medium with wavespeeds $c = 1$, $c = 2$, and $c = 4$. The results of trained WIDEBNET models on testing data are shown in Figure 13. We observe in Figure 13(b) that WIDEBNET manages to process the multiple arrivals to image the triangular scatterers. However, surprisingly, it does significantly poorer for the smoothly varying background. Explaining this discrepancy remains an open problem.

Remark: The notion of resolution becomes ambiguous for inhomogeneous media as the wavelength changes with background medium $c(x, z)$ following the dispersion

relation in (1.1). Nevertheless, across the range of background velocities the scatterers still contain sub-wavelength features such as e.g. the corners.

4.2.1. Comparison versus FWI. We compare WideBNet against FWI, implemented in Matlab, inverting for the same perturbation. The descent path is initialized with the known homogeneous background, and the gradient is computed using standard adjoint state methods. We selected as the optimization method L-BFGS implemented using the `fminunc` routine.

Following standard practices in the geophysical community we use a frequency sweep to regularize against the non-convexity of the objective function. We tested a dozen frequency combinations, and we selected the one which produced the best images. In the sweep, the data at different frequencies are fed to the optimization loop at three stages. At each stage we process data only at a certain frequency, without combining them, but we use the estimate at the end of one stage to initialize the subsequent stage one: in the first stage we process the lowest frequency data, we save the final answer which will be used as an initial guess for the next stage, which will process data in the immediately higher frequency-band, and we repeat until data at all frequencies are processed.

We ran the optimization until either the residual stagnated around 10^{-6}, or the norm of the gradient fell below 10^{-4}. In order to avoid the inverse crime we use a fourth order finite different stencil in the FWI formulation, in contrast to the data which was generated using a five-point second order stencil. For completeness, we also computed the regularized least-squares (LS) estimate using the far-field asymptotics in (2.9), using only the highest frequency data, and the least squares (LS) estimate using the finite difference discretization of the problem (with 9-point stencil) with wide-band data. After a laborious search for the best reconstruction we found that regularization parameter $\epsilon = 1$ for LS produced the best localisation while simultaneously minimizing oscillatory artifacts. The linear system was solved using `gmres` with a tolerance of 10^{-3}. In Figure 10 we can observe that for this specific class of scatters WideBNet outperforms all the other methods, and provides a sharper image of the perturbation with the correct amplitude (the far-field LS was re-scaled in this case).

We can observe that the reflectors are properly placed but the result from our neural network provides a better localization, sharper corners, with far fewer oscillatory artifacts. We point out that procuring these images for FWI was labour- and time-intensive. It took roughly a day to test all the different frequency sweeps and the full computation. The full computation took roughly one minute and a half in average for FWI, around two minutes for LS and half a minute for far-field LS. The experiments were carried in a 16-core workstation with an AMD 2950X CPU and 128 GB of RAM. In contrast, the training stage for WideBNet took in total 12 hours, and the inference takes a fraction of a second, running on an Nvidia GTX 1080Ti graphics card.

5. Conclusion & Future Work. In this manuscript we have designed an end-to-end architecture that is specifically tailored for solving the inverse scattering problem. We have shown that by assimilating multi-frequency data and coupling them through non-linearities we can produce images that solve the inverse scattering problem. Our tool produces results which are competitive with optimization-based approaches, but at a fraction of the cost. More critically, we have demonstrated that our architecture design and data assimilation strategy avoids three known shortcomings with conventional architectures and also other butterfly-based networks: (i) by incorporating tools from computational harmonic analysis, such as the butterfly fac-

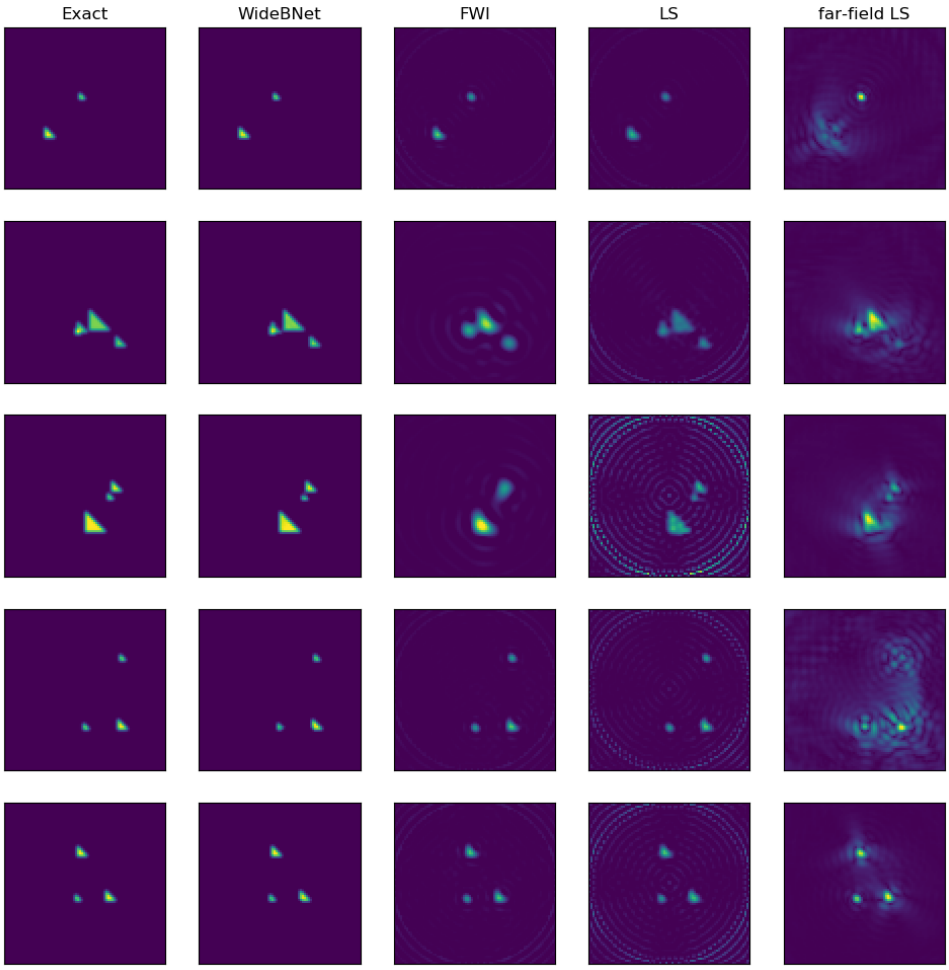

FIG. 10. *From left to right columns: exact perturbation, prediction by* WIDEBNET *, reconstruction by FWI using data at* $\{2.5Hz, 5.0Hz, 10.0Hz\}$*, least-squares using the finite difference modeling and data at* $\{2.5Hz, 5.0Hz, 10.0Hz\}$*, least-square using far-field approximation and data at* $\{10.0Hz\}$

torization, and multi-scale methods, such as the Cooley-Tukey FFT algorithm, we are able to drastically reduce the number of trainable parameters to match the inherent complexity of the problem and lower the training data requirements, (ii) our network has stable training dynamics and does not encounter issues such poorly conditioned gradients or poor local minima, and (iii) our network can be initialized using standard off-the-shelf technologies.

In addition, we have shown that our network recovers features below the diffraction limit of general, albeit fixed, class of scatterers. Even though there is an underlying assumption on the distribution of the scatterers we do not explicitly exploit it. Thus, one future research direction is to use the current architecture within a VAE or GAN framework, to fully capture the underlying distribution, and to further study the limits of the current architecture to image sub-wavelength features. Following the same approach one can seek to extend the applicability of the current architecture to the cases where there is noise in signal, or uncertainty on the background medium.

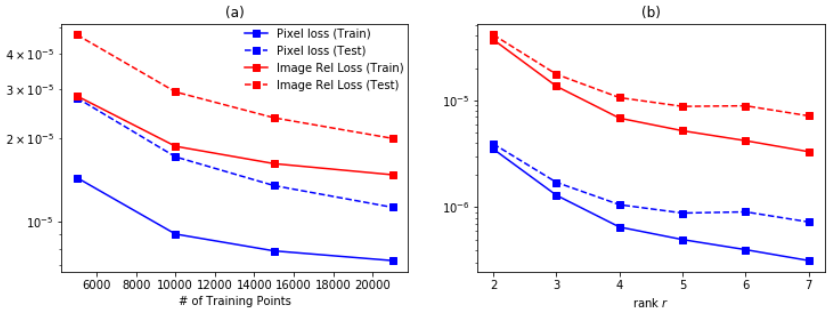

FIG. 11. *Sensitivity to hyper-parameters: number of training points and the rank* r. *(a) Performance of* WIDEBNET *with increasing number of training points. Note the same testing dataset, consisting of* 3000 *points, was used for all experiments. Observe that the generalization gap (the distance between the dashed and solid lines) remains asymptotically as both training and testing errors saturate. This exhibits that we are saturating the model capacity. (b) Performance of* WIDEBNET *with increasing rank* r. *The testing dataset was fixed for all experiments. Note that for leaf size* $s = 5$ *the maximum rank of the linearized model is* $r_{max} = 25$. *Although the training error decreases with increasing rank, we observe that the testing error begins to plateau beyond* $r = 3$.

Acknowledgments. We thank Yuehaw Khoo, Lexing Ying, Guillaume Bal, Yingzhou Li, Zhilong Fang, Pawan Bhawadraj, and Nori Nakata for fruitful discussions. We also thank George Barbastathis for detailed feedback on an earlier draft, and for invaluable references. In addition, we thank the two anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] M. ABADI, A. AGARWAL, P. BARHAM, E. BREVDO, Z. CHEN, C. CITRO, G. S. CORRADO, A. DAVIS, J. DEAN, M. DEVIN, S. GHEMAWAT, I. GOODFELLOW, A. HARP, G. IRVING, M. ISARD, Y. JIA, R. JOZEFOWICZ, L. KAISER, M. KUDLUR, J. LEVENBERG, D. MANÉ, R. MONGA, S. MOORE, D. MURRAY, C. OLAH, M. SCHUSTER, J. SHLENS, B. STEINER, I. SUTSKEVER, K. TALWAR, P. TUCKER, V. VANHOUCKE, V. VASUDEVAN, F. VIÉGAS, O. VINYALS, P. WARDEN, M. WATTENBERG, M. WICKE, Y. YU, AND X. ZHENG, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015, https://www.tensorflow.org/. Software available from tensorflow.org.

[2] H. K. AGGARWAL, M. P. MANI, AND M. JACOB, *MoDL: Model-based deep learning architecture for inverse problems*, IEEE Transactions on Medical Imaging, 38 (2019), pp. 394–405.

[3] T. ALKHALIFAH, *Scattering-angle based filtering of the waveform inversion gradients*, Geophys. J. Int., 200 (2014), pp. 363–373, https://doi.org/10.1093/gji/ggu379.

[4] D. ATKINSON AND N. D. APARICIO, *An inverse problem method for crack detection in viscoelastic materials under anti-plane strain*, Int. J. Eng. Sci., 35 (1997), pp. 841 – 849, https://doi.org/10.1016/S0020-7225(97)80003-1.

[5] G. BACKUS AND F. GILBERT, *The Resolving Power of Gross Earth Data*, Geophys. J. Int., 16 (1968), pp. 169–205, https://doi.org/10.1111/j.1365-246X.1968.tb00216.x.

[6] E. BAYSAL, D. D. KOSLOFF, AND J. W. C. SHERWOOD, *Reverse time migration*, GEOPHYSICS, 48 (1983), pp. 1514–1524, https://doi.org/10.1190/1.1441434.

[7] Y. BENGIO, P. SIMARD, AND P. FRASCONI, *Learning long-term dependencies with gradient descent is difficult*, IEEE Transactions on Neural Networks, 5 (1994), pp. 157–166, https://doi.org/10.1109/72.279181.

[8] J.-P. BÉRENGER, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[9] D. BLALOCK, J. J. G. ORTIZ, J. FRANKLE, AND J. GUTTAG, *What is the state of neural network pruning?*, 2020, https://arxiv.org/abs/2003.03033.

[10] C. BORGES, A. GILLMAN, AND L. GREENGARD, *High resolution inverse scattering in two di-*

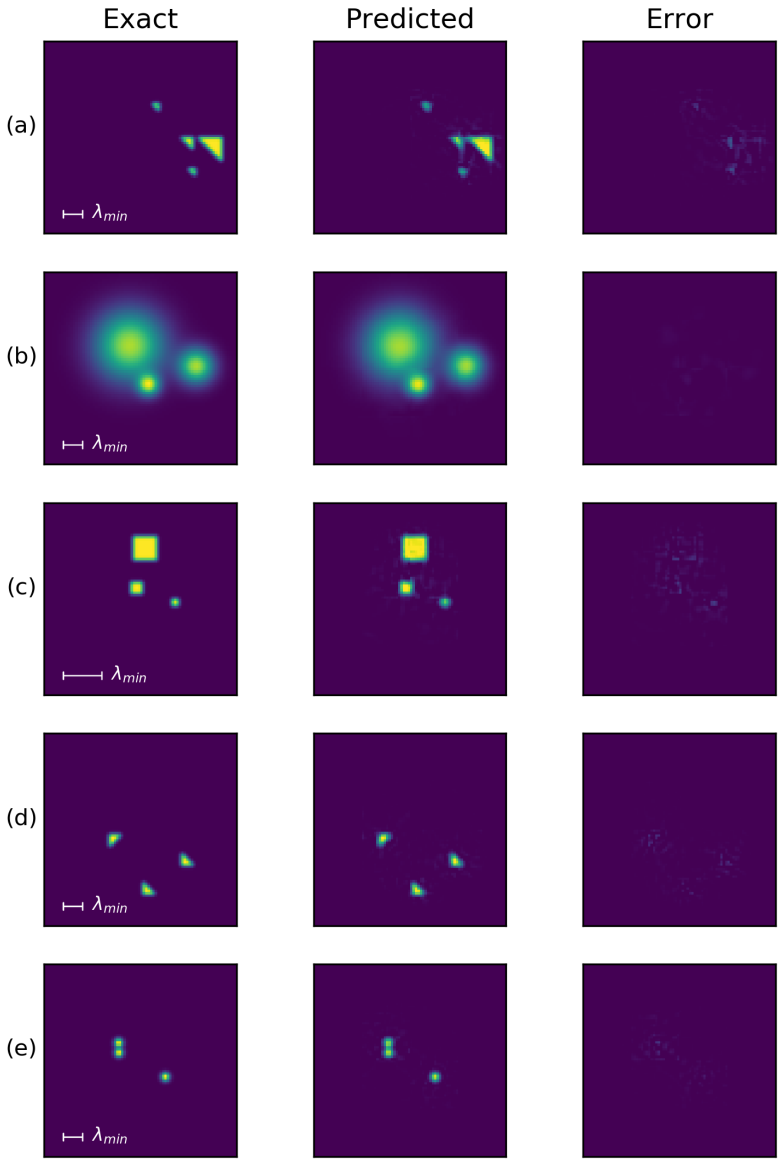

FIG. 12. *Visualization of* WIDEBNET *predictions on a testing set. The first column is the exact solution, the second column the output of* WIDEBNET *, and the third column the point-wise error. The colour scales in each row are normalized with respect to the first column. (a) with* $\triangle$ *(3,5,10). (b) same as above but with the Gaussian dataset. (c) heterogeneous squares but with a lower bandwidth (1.25, 2, and 5 Hz) so the effective wavelength is* 16 *PPW. (d) rotated dictionary. (e) Gaussian scatterers with characteristic length* 1.

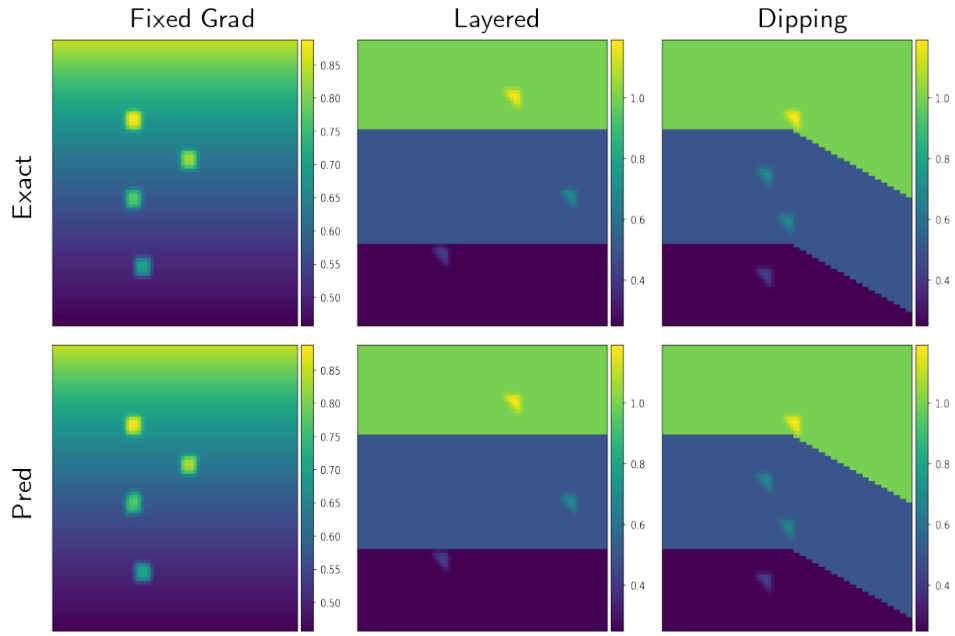

FIG. 13. *Visualization of* WIDEBNET *predictions on a testing set with inhomogeneous backgrounds. The colour scales of each row is normalized to the first column. The background medium is assumed known. (a) With a linearly increasing gradient in the background. (b) A layered medium increasing velocity with depth. (c) A layered medium with dipping reflectors. Note that the last configuration is not translation invariant.*

mensions using recursive linearization, SIAM J. Imaging Sci., 10 (2017), pp. 641–664, https://doi.org/10.1137/16M1093562.

[11] S. BÖRM, C. BÖRST, AND J. M. MELENK, *An analysis of a butterfly algorithm*, Comput. Math. Appl., 74 (2017), pp. 2125 – 2143, https://doi.org/10.1016/j.camwa.2017.05.019. Advances in Mathematics of Finite Elements, honoring 90th birthday of Ivo Babuška.

[12] J. BRUNA AND S. MALLAT, *Invariant scattering convolution networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 1872–1886.

[13] M. BURGER AND S. J. OSHER, *A survey on level set methods for inverse problems and optimal design*, Eur. J. Appl. Math., 16 (2005), p. 263–301, https://doi.org/10.1017/S0956792505006182.

[14] W. CAI, X. LI, AND L. LIU, *PhaseDNN - a parallel phase shift deep neural network for adaptive wideband learning*, 2019, https://arxiv.org/abs/1905.01389.

[15] W. CAI AND Z.-Q. J. XU, *Multi-scale deep neural networks for solving high dimensional PDEs*, ArXiv e-prints, [cs.LG] 1910.11710 (2019), https://arxiv.org/abs/1910.11710.

[16] E. CANDÈS, L. DEMANET, AND L. YING, *A fast butterfly algorithm for the computation of Fourier integral operators*, Multiscale Model. Sim., 7 (2009), pp. 1727–1750, https://doi.org/10.1137/080734339.

[17] E. J. CANDÈS AND C. FERNANDEZ-GRANDA, *Super-resolution from noisy data*, Journal of Fourier Analysis and Applications, 19 (2013), pp. 1229–1254, https://doi.org/10.1007/s00041-013-9292-3.

[18] E. J. CANDÈS AND C. FERNANDEZ-GRANDA, *Towards a mathematical theory of super-resolution*, Comm. Pure and Appl. Math., 67 (2014), pp. 906–956, https://doi.org/10.1002/cpa.21455.

[19] Y. CHEN, L. LU, G. E. KARNIADAKIS, AND L. D. NEGRO, *Physics-informed neural networks for inverse problems in nano-optics and metamaterials*, Opt. Express, 28 (2020), pp. 11618–11633, https://doi.org/10.1364/OE.384875.

[20] M. CHENEY, *A mathematical tutorial on synthetic aperture radar*, SIAM Rev., 43 (2001),

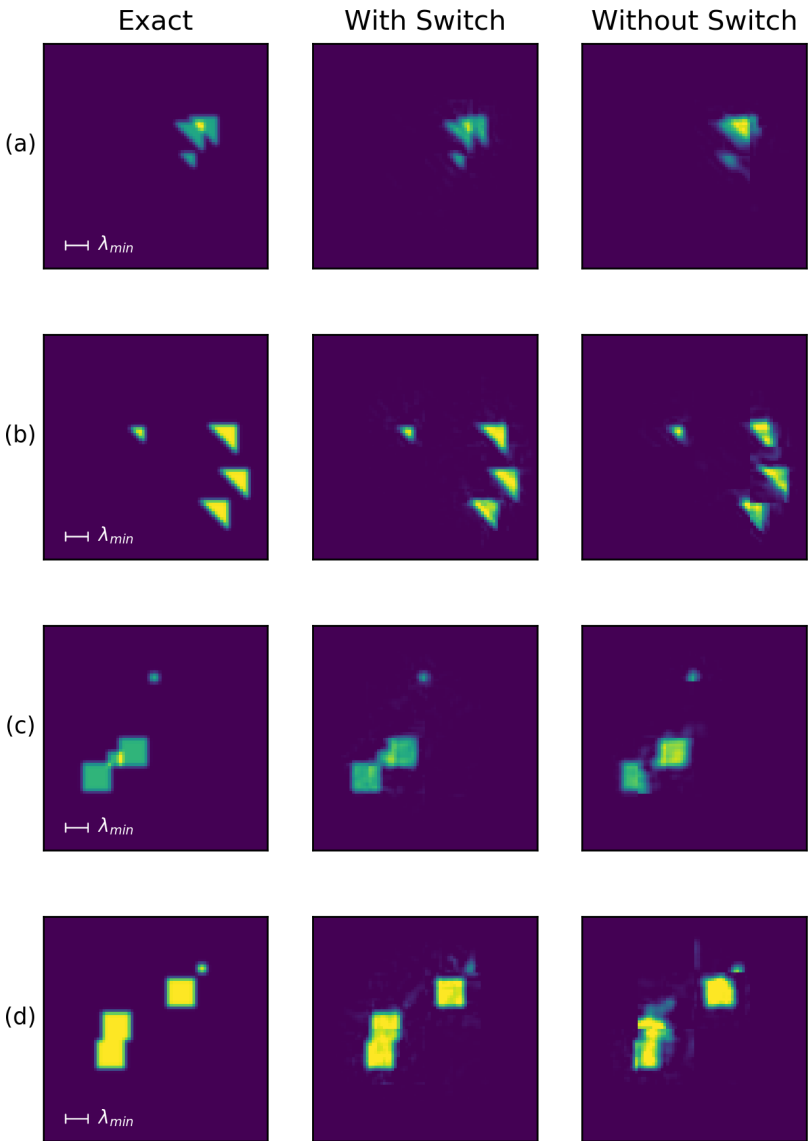

FIG. 14. *Visualization the effect of removing the switch permutation layer. The colour scale of each row is normalized to the first column. We observe that while* WIDEBNET *-without-switch manages to localize the scatterers, it is unable to fully resolve all sub-wavelength features.*

pp. 301–312, https://doi.org/10.1137/S0036144500368859.

[21] B. A. CIPRA, *The best of the 20th century: Editors name top 10 algorithms*, SIAM News, 33 (2000), pp. 1–2.

[22] N. COHEN, O. SHARIR, AND A. SHASHUA, *On the expressive power of deep learning: A tensor*

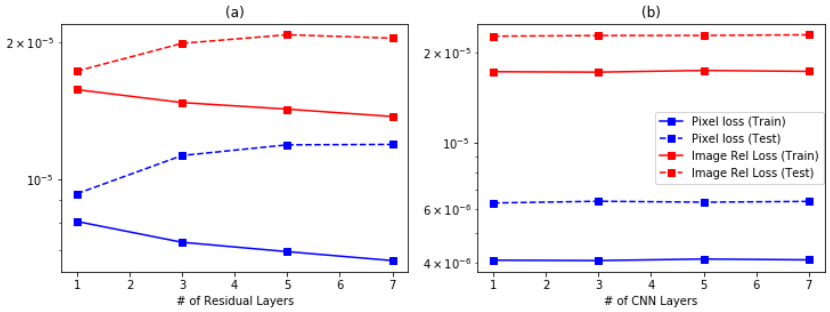

FIG. 15. *Sensitivity to hyper-parameters: number of residual layers and number of convolution post-processing layers. The testing set of* 3000 *points was fixed across all experiments. (a) The training error decreases with increasing residual layers, but the testing error increases. Note however the variation is negligible. (These experiments all had three convolution layers) . (b)* WIDEBNET *exhibits nearly complete insensitivity to the number of CNN post-processing layers. (this experiment was with three residual layers)*

analysis, in Conference on Learning Theory, 2016, pp. 698–728.

[23] A. COLLI, D. PRATI, M. FRAQUELLI, S. SEGATO, P. P. VESCOVI, F. COLOMBO, C. BALDUINI, S. DELLA VALLE, AND G. CASAZZA, *The use of a pocket-sized ultrasound device improves physical examination: Results of an in- and outpatient cohort study*, PLOS ONE, 10 (2015), pp. 1–10, https://doi.org/10.1371/journal.pone.0122181.

[24] D. COLTON AND A. KIRSCH, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393, https://doi.org/10.1088/0266-5611/12/4/003.

[25] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013, https://doi.org/10.1137/1.9781611973167.

[26] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag New York, New York, PA, 3 ed., 2013, https://doi.org/10.1007/978-1-4614-4942-3.

[27] J. W. COOLEY AND J. W. TUKEY, *An algorithm for the machine calculation of complex Fourier series*, Math. Comput., 19 (1965), pp. 297–301, http://www.jstor.org/stable/2003354.

[28] T. DAO, A. GU, M. EICHHORN, A. RUDRA, AND C. RÉ, *Learning fast algorithms for linear transforms using butterfly factorizations*, Proceedings of Machine Learning Research, 97 (2019), pp. 1517–1527.

[29] M. DE BUHAN AND M. DARBAS, *Numerical resolution of an electromagnetic inverse medium problem at fixed frequency*, Comput. Math. Appl., 74 (2017), pp. 3111 – 3128, https://doi.org/10.1016/j.camwa.2017.08.002.

[30] M. DE BUHAN AND M. KRAY, *A new approach to solve the inverse scattering problem for waves: combining the TRAC and the adaptive inversion methods*, Inverse Probl., 29 (2013), p. 085009, https://doi.org/10.1088/0266-5611/29/8/085009.

[31] L. DEMANET AND L. YING, *Fast wave computation via Fourier integral operators*, Math. Comput., 81 (2012), pp. 1455–1486.

[32] D. L. DONOHO, *Superresolution via sparsity constraints*, SIAM Journal on Mathematical Analysis, 23 (1992), pp. 1309–1331, https://doi.org/10.1137/0523074.

[33] B. ENGQUIST AND L. YING, *Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers*, Multiscale Model. Sim., 9 (2011), pp. 686–710.

[34] Y. FAN, J. FELIU-FABÀ, L. LIN, L. YING, AND L. ZEPEDA-NÚÑEZ, *A multiscale neural network based on hierarchical nested bases*, Research in the Mathematical Sciences, 6 (2019), p. 21, https://doi.org/10.1007/s40687-019-0183-3.

[35] Y. FAN AND L. YING, *Solving inverse wave scattering with deep learning*, arXiv:1911.13202, (2019).

[36] Y. FAN AND L. YING, *Solving optical tomography with deep learning*, arXiv:1910.04756, (2019).

[37] Y. FAN AND L. YING, *Solving traveltime tomography with deep learning*, arXiv:1911.11636, (2019).

[38] A. FICHTNER AND J. TRAMPERT, *Resolution analysis in full waveform inversion*, Geophysi-

cal Journal International, 187 (2011), pp. 1604–1624, https://doi.org/10.1111/j.1365-246x.2011.05218.x.

[39] J. Garnier and G. Papanicolaou, *Passive Imaging with Ambient Noise*, Cambridge Monographs on Applied and Computational Mathematic, Cambridge University Press, 2016, https://books.google.com/books?id=9kfGCwAAQBAJ.

[40] D. Gilton, G. Ongie, and R. Willett, *Neumann networks for inverse problems in imaging*, arXiv preprint arXiv:1901.03707, (2019).

[41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, 2016.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., Curran Associates, Inc., 2014, pp. 2672–2680, http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[43] B. Gutenberg, *Ueber Erdbebenwellen. VII A. Beobachtungen an Registrierungen von Fernbeben in Göttingen und Folgerung über die Konstitution des Erdkörpers (mit Tafel)*, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, 1914 (1914), pp. 125–176, http://eudml.org/doc/58907.

[44] K. He and J. Sun, *Convolutional neural networks at constrained time cost*, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015), pp. 5353–5360.

[45] L. Hörmander, *The Analysis of Linear Partial Differential Operators. IV: Fourier Integral Operators*, vol. 63 of Classics in Mathematics, Springer, Berlin, 2009.

[46] K. Hornik, M. Stinchcombe, and H. White, *Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks*, Neural networks, 3 (1990), pp. 551–560.

[47] P. Hähner and T. Hohage, *New stability estimates for the inverse acoustic inhomogeneous medium problem and applications*, SIAM Journal on Mathematical Analysis, 33 (2001), pp. 670–685, https://doi.org/10.1137/S0036141001383564.

[48] M. Innes, A. Edelman, K. Fischer, C. Rackauckas, E. Saba, V. B. Shah, and W. Tebbutt, *A differentiable programming system to bridge machine learning and scientific computing*, 2019, https://arxiv.org/abs/1907.07587.

[49] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 5967–5976, https://doi.org/10.1109/CVPR.2017.632.

[50] E. Kang, W. Chang, J. Yoo, and J. C. Ye, *Deep convolutional framelet denosing for low-dose CT via wavelet residual network*, IEEE Transactions on Medical Imaging, 37 (2018), pp. 1358–1369.

[51] Y. Khoo and L. Ying, *SwitchNet: A neural network model for forward and inverse scattering problems*, SIAM J. Sci. Comput., 41 (2019), pp. A3182–A3201, https://doi.org/10.1137/18M1222399.

[52] D. Kingma and J. Ba, *Adam: a method for stochastic optimization*, in Proceedings of the International Conference on Learning Representations (ICLR), May 2015.

[53] D. P. Kingma and M. Welling, *Auto-encoding variational Bayes*, 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, (2014), pp. 1–14, https://arxiv.org/abs/1312.6114.

[54] A. Kirsch and N. Grinberg, *The Factorization Method for Inverse Problems*, Oxford University Press, Oxford, first ed., 2008.

[55] K. Kothari, M. de Hoop, and I. Dokmanić, *Learning the geometry of wave-based imaging*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 8318–8329, https://proceedings.neurips.cc/paper/2020/file/5e98d23afe19a774d1b2dcbefd5103eb-Paper.pdf.

[56] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), p. 436.

[57] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, *Photo-realistic single image super-resolution using a generative adversarial network*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 105–114, https://doi.org/10.1109/CVPR.2017.19.

[58] Y. Li, X. Cheng, and J. Lu, *Butterfly-Net: Optimal function representation based on convolutional neural networks*, arXiv preprint arXiv:1805.07451, (2018).

[59] Y. Li, H. Yang, E. Martin, K. Ho, and L. Ying, *Butterfly factorization*, Multiscale Model. Sim., 13 (2015), pp. 714–732, https://doi.org/10.1137/15M1007173.

[60] Y. Li, H. Yang, and L. Ying, *Multidimensional butterfly factorization*, Applied and Compu-

tational Harmonic Analysis, 44 (2018), pp. 737 – 758, https://doi.org/https://doi.org/10.1016/j.acha.2017.04.002.

[61] Y. E. Li and L. Demanet, *Full-waveform inversion with extrapolated low-frequency data*, GEOPHYSICS, 81 (2016), pp. R339–R348, https://doi.org/10.1190/geo2016-0038.1.

[62] Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, *Fourier Neural Operator for Parametric Partial Differential Equations*, in International Conference on Learning Representations, 2021, https://openreview.net/forum?id=c8P9NQVtmnO.

[63] Y. Liu, X. Xing, H. Guo, E. Michielssen, and X. S. Ghysels, P. Li, *Butterfly factorization via randomized matrix-vector multiplications*, arXiv:2002.03400, (2020).

[64] T. Luo, Z. Ma, Z.-Q. J. Xu, and Y. Zhang, *Theory of the frequency principle for general deep neural networks*, ArXiv e-prints, [cs.LG] 1906.09235 (2018), https://arxiv.org/abs/arXiv:1906.09235.

[65] X. Mao, C. Shen, and Y.-B. Yang, *Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections*, in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Inc., 2016, pp. 2802–2810.

[66] H. Mhaskar, Q. Liao, and T. Poggio, *Learning functions: When is deep better than shallow*, arXiv preprint arXiv:1603.00988, (2016).

[67] M. Mirza and S. Osindero, *Conditional generative adversarial nets*, 2014, http://arxiv.org/abs/1411.1784.

[68] M. Oberguggenberger and M. Schwarz, *Wave propagation in random media, parameter estimation and damage detection via stochastic Fourier integral operators*, arXiv:2009.09389, (2020).

[69] R. D. Oldham, *The constitution of the interior of the Earth, as revealed by earthquakes*, Quarterly Journal of the Geological Society, 62 (1906), pp. 456–475, https://doi.org/10.1144/GSL.JGS.1906.062.01-04.21.

[70] M. O'Neil, F. Woolfe, and V. Rokhlin, *An algorithm for the rapid evaluation of special function transforms*, Appl. Comput. Harmon. A., 28 (2010), pp. 203 – 226, https://doi.org/10.1016/j.acha.2009.08.005. Special Issue on Continuous Wavelet Transform in Memory of Jean Morlet, Part I.

[71] S. Pang and G. Barbastathis, *Machine Learning Regularized Solution of the Lippmann-Schwinger Equation*, arXiv:2010.15117, (2020).

[72] J. R. Pettit, A. E. Walker, and M. J. S. Lowe, *Improved detection of rough defects for ultrasonic nondestructive evaluation inspections based on finite element modeling of elastic wave scattering*, IEEE T. Ultrason. Ferr., 62 (2015), pp. 1797–1808, https://doi.org/10.1109/TUFFC.2015.007140.

[73] J. Poulson, L. Demanet, N. Maxwell, and L. Ying, *A parallel butterfly algorithm*, SIAM J. Sci. Comput., 36 (2014), pp. C49–C65, https://doi.org/10.1137/130921544.

[74] R. G. Pratt, *Seismic waveform inversion in the frequency domain; part 1: Theory and verification in a physical scale model*, GEOPHYSICS, 64 (1999), pp. 888–901, https://doi.org/10.1190/1.1444597.

[75] M. Raissi and G. E. Karniadakis, *Hidden physics models: Machine learning of nonlinear partial differential equations*, J. Comput. Phys., 357 (2018), pp. 125 – 141, https://doi.org/10.1016/j.jcp.2017.11.039, http://www.sciencedirect.com/science/article/pii/S0021999117309014.

[76] N. Rawlinson, S. Pozgay, and S. Fishwick, *Seismic tomography: A window into deep Earth*, Phys. Earth Planet. Int., 178 (2010), pp. 101–135, https://doi.org/10.1016/j.pepi.2009.10.002.

[77] D. J. Rezende, S. Mohamed, and D. Wierstra, *Stochastic backpropagation and approximate inference in deep generative models*, in International Conference on Machine Learning, PMLR, Jun. 2014, pp. 1278–1286, http://proceedings.mlr.press/v32/rezende14.html.

[78] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Springer International Publishing, Cham, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[79] H. Schomberg, *An improved approach to reconstructive ultrasound tomography*, J. of Phys. D: Appl. Phys., 11 (1978), pp. L181–L185, https://doi.org/10.1088/0022-3727/11/15/004.

[80] P. Stefanov, G. Uhlmann, A. Vasy, and H. Zhou, *Travel time tomography*, Acta Math. Sin., 35 (2019), pp. 1085–1114, https://doi.org/10.1007/s10114-019-8338-0.

[81] W. W. Symes and J. J. Carazzone, *Velocity inversion by differential semblance optimization*, GEOPHYSICS, 56 (1991), pp. 654–663, https://doi.org/10.1190/1.1443082.

[82] A. Tarantola, *Inversion of seismic reflection data in the acoustic approximation*, GEO-

PHYSICS, 49 (1984), pp. 1259–1266, https://doi.org/10.1190/1.1441754.

[83] T. van Leeuwen and F. J. Herrmann, *Mitigating local minima in full-waveform inversion by expanding the search space*, Geophys. J. Int., 195 (2013), pp. 661–667, https://doi.org/10.1093/gji/ggt258.

[84] J. Virieux, A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou, *6. An introduction to full waveform inversion*, Society of Exploration Geophysicists, 2017, pp. R1–1–R1–40, https://doi.org/10.1190/1.9781560803027.entry6, https://library.seg.org/doi/abs/10.1190/1.9781560803027.entry6, https://arxiv.org/abs/https://library.seg.org/doi/pdf/10.1190/1.9781560803027.entry6.

[85] J. Virieux and S. Operto, *An overview of full-waveform inversion in exploration geophysics*, GEOPHYSICS, 74 (2009), pp. WCC1–WCC26, https://doi.org/10.1190/1.3238367.

[86] Z. Xu, Y. Li, and X. Cheng, *Butterfly-Net2: Simplified Butterfly-Net and Fourier transform initialization*, in Proceedings of The First Mathematical and Scientific Machine Learning Conference, J. Lu and R. Ward, eds., vol. 107 of Proceedings of Machine Learning Research, Princeton University, Princeton, NJ, USA, 20–24 Jul. 2020, PMLR, pp. 431–450, http://proceedings.mlr.press/v107/xu20b.html.

[87] Z.-Q. J. Xu, *Frequency principle in deep learning with general loss functions and its potential application*, ArXiv e-prints, [cs.LG] 1811.10146 (2018), https://arxiv.org/abs/arXiv:1811.10146.

[88] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, *Training behavior of deep neural network in frequency domain*, in Neural Information Processing, T. Gedeon, K. W. Wong, and M. Lee, eds., Cham, 2019, Springer International Publishing, pp. 264–274.

[89] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, *DeepSense: A unified deep learning framework for time-series mobile sensing data processing*, in Proceedings of the 26th International Conference on World Wide Web, WWW '17, Republic and Canton of Geneva, CHE, 2017, International World Wide Web Conferences Steering Committee, p. 351–360, https://doi.org/10.1145/3038912.3052577.

[90] S. Yao, A. Piao, W. Jiang, Y. Zhao, H. Shao, S. Liu, D. Liu, J. Li, T. Wang, S. Hu, L. Su, J. Han, and T. Abdelzaher, *STFNets: Learning sensing signals from the time-frequency perspective with short-time Fourier neural networks*, in The World Wide Web Conference, WWW '19, New York, NY, USA, 2019, Association for Computing Machinery, p. 2192–2202, https://doi.org/10.1145/3308558.3313426.

[91] J. C. Ye, Y. Han, and E. Cha, *Deep convolutional framelets: A general deep learning framework for inverse problems*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 991–1048, https://doi.org/10.1137/17M1141771.

[92] L. Zepeda-Núñez, Y. Chen, J. Zhang, W. Jia, L. Zhang, and L. Lin, *Deep Density: circumventing the Kohn-sham equations via symmetry preserving neural networks*. https://www.math.wisc.edu/~lzepeda/Deep-Density.pdf, 2019.

[93] L. Zepeda-Núñez and L. Demanet, *The method of polarized traces for the 2D Helmholtz equation*, J. Comput. Phys., 308 (2016), pp. 347 – 388, https://doi.org/http://dx.doi.org/10.1016/j.jcp.2015.11.040.

[94] J. Zhang, L. Zepeda-Núñez, Y. Yao, and L. Lin, *Learning the mapping* $\mathbf{x} \mapsto \sum_{i=1}^{d} x_i^2$*: the cost of finding the needle in a haystack*, Comm. App. Math. Comp., (2020).

[95] L. Zhang, J. Han, H. Wang, R. Car, and W. E, *Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics*, Physical Review Letters, 120 (2018), p. 143001.

[96] L. Zhang, J. Han, H. Wang, R. Car, and W. E, *DeepCG: Constructing coarse-grained models via deep neural networks*, J. Chem. Phys., 149 (2018), p. 034101, https://doi.org/10.1063/1.5027645.

Appendix A. Permutation and Switch Indices.

The permutation indexing that enables the G^ℓ and H^ℓ layers to be described as block-diagonal matrices can be derived from an involved analysis of the two-dimensional butterfly factorization. For convenience, we provide a generic pseudocode in Alg. 6 which automates this construction for all choices of L. We note that this permutation assumes the input vector is Morton-flattened. Similarly, Alg. 7 yields pseudocode for generating the specific switch permutation indices of the Butterfly algorithm.

```
def permutation_indices(L, ℓ):
  # indices inside each 4^{L-ℓ} × 4^{L-ℓ} block
  Δ = 4^{L-ℓ-1}

  # [0,Δ,2Δ,3Δ,0,Δ,2Δ,3Δ,4Δ,...]
  π^ℓ = np.file(np.arange(4)*Δ, Δ)

  # + [0,0,0,0,1,1,1,1,...,Δ, Δ,Δ,Δ]
  π^ℓ += np.repeat(np.arange(Δ), 4)

  # indices for entire block diagonal matrix
  π^ℓ = np.tile(π^ℓ, 4^ℓ)
  π^ℓ += np.repeat(np.arange(4^ℓ)*4^{L-ℓ}, 4^{L-ℓ})

  return π^ℓ
```

Listing 6

Pseudo code for permutation pattern π_ℓ used in layers G^ℓ and H^ℓ when processing Morton-flattened inputs.

```
def switch_indices(L):
  π_switch = np.arange(2^L)*(2^L)

  π_switch = np.tile(π_switch, 2^L)
  π_switch += np.repeat(np.arange(2^L), 2^L)

  return π_switch
```

Listing 7

Pseudo code for switch permutation indices for processing Morton-flattened inputs.