TRANSFORMER MEETS BOUNDARY VALUE INVERSE PROBLEMS

Ruchi Guo
Department of Mathematics
University of California, Irvine

Shuhao Cao
Division of Computing, Analytics, and Mathematics
School of Science and Engineering
University of Missouri-Kansas City

Long Chen
Department of Mathematics
University of California, Irvine

ABSTRACT

A Transformer-based deep direct sampling method is proposed for electrical impedance tomography, a well-known severely ill-posed nonlinear boundary value inverse problem. A real-time reconstruction is achieved by evaluating the learned inverse operator between carefully designed data and the reconstructed images. An effort is made to give a specific example to a fundamental question: whether and how one can benefit from the theoretical structure of a mathematical problem to develop task-oriented and structure-conforming deep neural networks? Specifically, inspired by direct sampling methods for inverse problems, the 1D boundary data in different frequencies are preprocessed by a partial differential equation-based feature map to yield 2D harmonic extensions as different input channels. Then, by introducing learnable non-local kernels, the direct sampling is recast to a modified attention mechanism. The new method achieves superior accuracy over its predecessors and contemporary operator learners and shows robustness to noises in benchmarks. This research shall strengthen the insights that, despite being invented for natural language processing tasks, the attention mechanism offers great flexibility to be modified in conformity with the a priori mathematical knowledge, which ultimately leads to the design of more physics-compatible neural architectures.

1 INTRODUCTION

Boundary value inverse problems aim to recover the internal structure or distribution of multiple media inside an object (2D reconstruction) based on only the data available on the boundary (1D signal input), which arise from many imaging techniques, e.g., electrical impedance tomography (EIT) (Holder, 2004), diffuse optical tomography (DOT) (Culver et al., 2003), magnetic induction tomography (MIT) (Griffiths et al., 1999). Not needing any internal data renders these techniques generally non-invasive, safe, cheap, and thus quite suitable for monitoring applications.

In this work, we shall take EIT as an example to illustrate how a more structure-conforming neural network architecture leads to better results in certain physics-based tasks. Given a 2D bounded domain Ω and an inclusion D, the forward model is the following partial differential equation (PDE)

$$\nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } \Omega, \qquad \text{where } \ \sigma = \sigma_1 \text{ in } D, \text{ and } \ \sigma = \sigma_0 \text{ in } \Omega \backslash \overline{D}, \tag{1}$$

where σ is a piecewise constant function defined on Ω with known function values σ_0 and σ_1, but the shape of the inclusion D buried in Ω is unknown. The goal is to recover the shape of D using only the boundary data on $\partial\Omega$ (Figure 1). Specifically, by exerting a current g on the boundary, one solves (1) with the Neumann boundary condition $\sigma \nabla u \cdot \mathbf{n}|_{\partial\Omega} = g$, where $\mathbf{n}$ is the outwards unit normal direction of $\partial\Omega$, to get a unique u on the whole domain Ω. In practice, only the Dirichlet boundary value representing the voltages $f = u|_{\partial\Omega}$ on the boundary can be measured. This procedure is called Neumann-to-Dirichlet (NtD) mapping:

$$\Lambda_\sigma : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega), \quad \text{with } \ g = \sigma \nabla u \cdot \mathbf{n}|_{\partial\Omega} \mapsto f = u|_{\partial\Omega}. \tag{2}$$

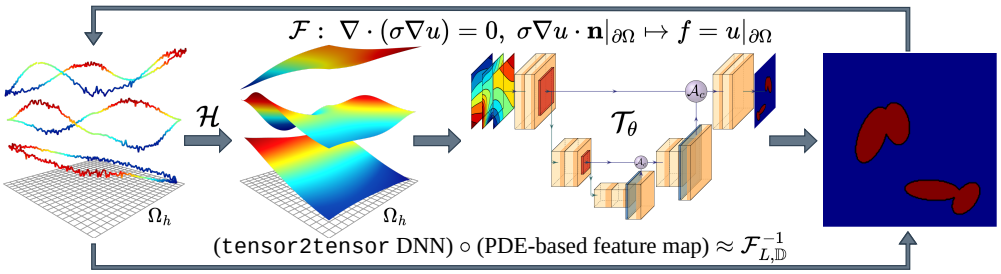

Figure 1: Schematics of the pipeline: the approximation of the inverse operator is decomposed into (i) a PDE-based feature map $\mathcal{H}$: the 1D boundary data is extended to 2D features by a harmonic extension; (ii) a tensor2tensor neural network $\mathcal{T}_\theta$ that outputs the reconstruction.

For various notation and the Sobolev space formalism, we refer readers to Appendix A; for a brief review of the theoretical background of EIT we refer readers to Appendix B. The NtD map above in (2) can be expressed as

$$\mathbf{f} = \mathbf{A}_\sigma \mathbf{g}, \tag{3}$$

where $\mathbf{g}$ and $\mathbf{f}$ are (infinite-dimensional) vector representations of functions g and f relative to a chosen basis, and $\mathbf{A}_\sigma$ is the matrix representation of Λ_σ (see Appendix B for an example).

The original mathematical setup of EIT is to use the NtD map Λ_σ in (2) to recover σ, referred to as the case of full measurement (Calderón, 2006). In this case, the forward and inverse operators associated with EIT can be formulated as

$$\mathcal{F} \ : \ \sigma \mapsto \Lambda_\sigma, \quad \text{and} \quad \mathcal{F}^{-1} \ : \ \Lambda_\sigma \mapsto \sigma. \tag{4}$$

Fix a set of basis $\{g_l\}_{l=1}^\infty$ of the corresponding Hilbert space containing all admissible currents. Then, mathematically speaking, "knowing the operator Λ_σ" means that one can measure all the current-to-voltage pairs $\{g_l, f_l := \Lambda_\sigma g_l\}_{l=1}^\infty$ and construct the infinite-dimensional matrix $\mathbf{A}_\sigma$.

However, as infinitely many boundary data pairs are not attainable in practice, the problem of more practical interest is to use only a few data pairs $\{(g_l, f_l)\}_{l=1}^L$ for reconstruction. In this case, the forward and inverse problems can be formulated as

$$\mathcal{F}_L : \sigma \mapsto \{(g_1, \Lambda_\sigma g_1), ..., (g_L, \Lambda_\sigma g_L)\} \quad \text{and} \quad \mathcal{F}_L^{-1} : \{(g_1, \Lambda_\sigma g_1), ..., (g_L, \Lambda_\sigma g_L)\} \mapsto \sigma. \tag{5}$$

For limited data pairs, the inverse operator $\mathcal{F}_L^{-1}$ is extremely ill-posed or even not well-defined (Isakov & Powell, 1990; Barceló et al., 1994; Kang & Seo, 2001; Lionheart, 2004); namely, the same boundary measurements may correspond to different σ. In view of the matrix representation $\mathbf{A}_\sigma$, for $\mathbf{g}_l = \mathbf{e}_l$, $l = 1, ..., L$, with $\mathbf{e}_l$ being unit vectors of a chosen basis, $(\mathbf{f}_1, ..., \mathbf{f}_L)$ only gives the first L columns of $\mathbf{A}_\sigma$. It is possible that two matrices $\mathbf{A}_\sigma$ and $\mathbf{A}_{\tilde{\sigma}}$ have similar first L columns but $\|\sigma - \tilde{\sigma}\|$ is large.

How to deal with this "ill-posedness" is a central theme in boundary value inverse problem theories. The operator learning approach has the potential to tame the ill-posedness by restricting $\mathcal{F}_L^{-1}$ at a set of sampled data $\mathbb{D} := \{\sigma^{(k)}\}_{k=1}^N$, with different shapes and locations following certain distribution. Then the problem becomes to approximate

$$\mathcal{F}_{L,\mathbb{D}}^{-1} : \{(g_1, \Lambda_{\sigma^{(k)}} g_1), ..., (g_L, \Lambda_{\sigma^{(k)}} g_L)\} \mapsto \sigma^{(k)}, \quad k = 1, \ldots, N. \tag{6}$$

The fundamental assumption here is that this map is "well-defined" enough to be regarded as a high-dimensional interpolation (learning) problem on a compact data submanifold (Seo et al., 2019; Ghattas & Willcox, 2021), and the learned approximate mapping can be evaluated at newly incoming σ's. The incomplete information of Λ_σ due to a small L for one single σ is compensated by a large $N \gg 1$ sampling of different σ's.

2 Background, Related Work, and Contributions

Classical iterative methods. There are in general two types of methodology to solve inverse problems. The first one is a large family of iterative or optimization-based methods (Dobson &

Santosa, 1994; Martin & Idier, 1997; Chan & Tai, 2003; Vauhkonen et al., 1999; Guo et al., 2019; Rondi & Santosa, 2001; Chen et al., 2020; Bao et al., 2020; Gu et al., 2021). One usually looks for an approximated σ by solving a minimization problem with a regularization $\mathcal{R}(\sigma)$ to alleviate the ill-posedness, say

$$\inf_\sigma \left\{ \sum_{l=1}^{L} \|\Lambda_\sigma g_l - f_l\|_{\partial\Omega}^2 + \mathcal{R}(\sigma) \right\}. \tag{7}$$

The design of regularization $\mathcal{R}(\sigma)$ plays a critical role in a successful reconstruction (Tarvainen et al., 2008; Tehrani et al., 2012; Wang et al., 2012). Due to the ill-posedness, the computation for almost all iterative methods usually takes numerous iterations to converge, and the reconstruction is highly sensitive to noise. Besides, the forward operator $\mathcal{F}(\cdot)$ needs to be evaluated at each iteration, which is itself expensive as it requires solving forward PDE models.

Classical direct methods. The second methodology is to develop a well-defined mapping $\mathcal{G}_\theta$ parametrized by θ, empirically constructed to approximate the inverse map itself, say $\mathcal{G}_\theta \approx \mathcal{F}^{-1}$. These methods are referred to as non-iterative or direct methods in the literature. Distinguished from iterative approaches, direct methods are typically highly problem-specific, as they are designed based on specific mathematical structures of their respective inverse operators. For instance, methods in EIT and DOT include factorization methods (Kirsch & Grinberg, 2007; Azzouz et al., 2007; Brühl, 2001; Hanke & Brühl, 2003), MUSIC-type algorithms (Cheney, 2001; Ammari & Kang, 2004; 2007; Lee et al., 2011), and the D-bar methods (Knudsen et al., 2007; 2009) based on a Fredholm integral equation (Nachman, 1996), among which are the direct sampling methods (DSM) being our focus in this work (Chow et al., 2014; 2015; Kang et al., 2018; Ito et al., 2013; Ji et al., 2019; Harris & Kleefeld, 2019; Ahn et al., 2020). These methods generally have a closed-form $\mathcal{G}_\theta$ for approximation, and the parameters θ represent model-specific mathematical objects. For each fixed θ, this procedure is usually much more stable than iterative approaches with respect to the input data. Furthermore, the evaluation for each boundary data pair is distinctly fast, as no optimization is needed. However, a simple closed-form $\mathcal{G}_\theta$ admitting efficient execution may not be available in practice since some mathematical assumptions and derivation may not hold. For instance, MUSIC-type and D-bar methods generally require an accurate approximation to Λ_σ, while DSM poses restrictions on the boundary data, domain geometry, etc., see Appendix D for details.

Boundary value inverse problems. For most cases of boundary value inverse problems in 2D, the major difference, e.g., with an inverse problem in computer vision (Marroquin et al., 1987), is that data are only available on 1D manifolds, which are used to reconstruct 2D targets. When comparing (4) with a linear inverse problem in signal processing $\mathbf{y} = A\mathbf{x} + \boldsymbol{\epsilon}$, to recover a signal $\mathbf{x}$ from measurement $\mathbf{y}$ with noise $\boldsymbol{\epsilon}$, the difference is more fundamental in that $\mathcal{F}(\cdot)$ itself is highly nonlinear and involves boundary value PDEs. Moreover, the boundary data themselves generally involve certain input-output structures (NtD maps), which adds more complexity. In Adler & Guardo (1994); Fernández-Fuentes et al. (2018); Feng et al. (2018), boundary measurements are collected and directly input into feedforward fully connected networks. As the data reside on different manifolds, special treatments are made to the input data, such as employing pre-reconstruction stages to generate rough 2D input to CNNs (Ben Yedder et al., 2018; Ren et al., 2020; Pakravan et al., 2021).

Deep neural network and inverse problems. Solving an inverse problem is essentially to give a satisfactory approximation to $\mathcal{F}^{-1}$ but based on finitely many measurements. The emerging deep learning (DL) based on Deep Neural Networks (DNN) to directly emulate operators significantly resembles those classical direct methods mentioned above. However, operator learners by DNNs are commonly considered black boxes. A natural question is how the a priori mathematical knowledge can be exploited to design more physics-compatible DNN architectures. In pursuing the answer to this question, we aim to provide a supportive example that bridges deep learning techniques and classical direct methods, which improves the reconstruction of EIT.

Operator learners. Operator learning has become an active research field for inverse problems in recent years, especially related to image reconstruction where CNN plays a central role, see, e.g., Kłosowski & Rymarczyk (2017); Nguyen et al. (2018); Tan et al. (2018); Jin et al. (2017); Kang et al. (2017); Barbastathis et al. (2019); Latif et al. (2019); Zhu et al. (2018); Chen et al. (2021); Coxson et al. (2022); Zhu et al. (2023b). Notable examples of efforts to couple classical reconstruction methods and CNN include Hamilton et al. (2019); Hamilton & Hauptmann (2018), where a CNN post-processes images obtained by the classical D-bar methods, and Fan et al. (2019); Fan & Ying (2020) where BCR-Net is developed to mimic pseudo-differential operators appearing in many inverse

problems. A deep direct sampling method is proposed in Guo & Jiang (2020); Guo et al. (2021) that learns local convolutional kernels mimicking the gradient operator of DSM. Another example is radial basis function neural networks seen in Hrabuska et al. (2018); Michalikova et al. (2014); Wang et al. (2021a). Nevertheless, convolutions in CNNs use kernels whose receptive fields involve only a small neighborhood of a pixel. Thus, layer-wise, CNN does not align well with the non-local nature of inverse problems. More recently, the learning of PDE-related forward problems using global kernel has gained attraction, most notably the Fourier Neural Operator (FNO) (Li et al., 2021a; Kovachki et al., 2021; Guibas et al., 2022; Zhao et al., 2022; Wen et al., 2022; Li et al., 2022b). FNO takes advantage of the low-rank nature of certain problems and proposes to train a local kernel in the frequency domain yet global in the spatial-temporal domain, mimicking the solution's kernel integral form. Concurrent studies include DeepONets (Lu et al., 2021; Wang et al., 2021b; Jin et al., 2022b), Transformers (Cao, 2021; Kissas et al., 2022; Li et al., 2022a; Liu et al., 2022; Fonseca et al., 2023), Integral Autoencoder (Ong et al., 2022), Multiwavelet Neural Operators (Gupta et al., 2021; 2022), and others (Lütjens et al., 2022; Hu et al., 2022; Boussif et al., 2022; de Hoop et al., 2022a;b; Ryck & Mishra, 2022; Seidman et al., 2022; Zhang et al., 2023; Lee, 2023; Zhu et al., 2023a).

Related studies on Transformers. The attention mechanism-based models have become state of the art in many areas since Vaswani et al. (2017). One of the most important and attractive aspects of the attention mechanism is its unparalleled capability to efficiently model non-local long-range interactions (Katharopoulos et al., 2020; Choromanski et al., 2021; Nguyen et al., 2021). The relation of the attention with kernel learning is first studied in Tsai et al. (2019) and later connected with random feature (Peng et al., 2021). Connecting the non-PDE-based integral kernels and the attention mechanism has been seen in Hutchinson et al. (2021); Guibas et al. (2022); Nguyen et al. (2022); Han et al. (2022). Among inverse problems, Transformers have been applied in medical imaging applications, including segmentation (Zhou et al., 2021; Hatamizadeh et al., 2022; Petit et al., 2021), X-Ray (Tanzi et al., 2022), magnetic resonance imaging (MRI) (He et al., 2022), ultrasound (Perera et al., 2021), optical coherence tomography (OCT) (Song et al., 2021). To our best knowledge, no work in the literature establishes an architectural connection between the attention mechanism in Transformer and the mathematical structure of PDE-based inverse problems.

2.1 Contributions

- *A structure-conforming network architecture.* Inspired by the EIT theory and classic DSM, we decompose the approximation of the inverse operator into a harmonic extension and an integral operator with learnable non-local kernels that has an attention-like structure. Additionally, the attention architecture is reinterpreted through a Fredholm integral operator to rationalize the application of the Transformer to the boundary value inverse problem.
- *Theoretical and experimental justification for the advantage of Transformer.* We have proved that, in Transformers, modified attention can represent target functions exhibiting higher frequency natures from lower frequency input features. A comparative study in the experiments demonstrates a favorable match between the Transformer and the benchmark problem.

3 Interplay between mathematics and neural architectures

In this section, we try to articulate that the triple tensor product in the attention mechanism matches exceptionally well with representing a solution in the inverse operator theory of EIT. In pursuing this end goal, this study tries to answer the following motivating questions:

(**Q1**) What is an appropriate finite-dimensional data format as inputs to the neural network?

(**Q2**) Is there a suitable neural network matching the mathematical structure?

3.1 From EIT to Operator Learning

In the case of full measurement, the operator $\mathcal{F}^{-1}$ can be well approximated through a large number of $(\sigma, \mathbf{A}_\sigma)$ data pairs. This mechanism essentially results in a *tensor2tensor* mapping/operator from $\mathbf{A}_\sigma$ to the imagery data representing σ. In particular, the BCR-Net (Fan & Ying, 2020) is a DNN approximation falling into this category. However, when there are very limited boundary data pairs accessible, the task of learning the full matrix $\mathbf{A}_\sigma$ becomes obscure, which complicates the development of a *tensor2tensor* pipeline for operator learning.

Operator learning problems for EIT. We first introduce several attainable approximations of infinite-dimensional spaces by finite-dimensional counterparts for the proposed method.

(1) *Spatial discretization.* Let Ω_h be a mesh of Ω with the mesh spacing h and let $\{z_j\}_{j=1}^M := \mathcal{M}$ be the set of grid points to represent the 2D discretization of continuous signals. Then a function u defined almost everywhere in Ω can be approximated by a vector $\mathbf{u}_h \in \mathbb{R}^M$.

(2) *Sampling of D.* We generate N samples of D with different shapes and locations following certain distributions. For example, elliptical inclusions with random semi-axes and centers are generated as a benchmark (see Appendix C.1 for details). With the known σ_0 and σ_1, set the corresponding data set $\mathbb{D} = \{\sigma^{(1)}, \sigma^{(2)}, ..., \sigma^{(N)}\}$. N is usually large enough to represent field applications of interest.

(3) *Sampling of NtD maps.* For the k-th sample of D, we generate L pairs of boundary data $\{(g_l^{(k)}, f_l^{(k)})\}_{l=1}^L$ by solving PDE (1), which can be thought of as sampling of columns of the infinite matrix $\mathbf{A}_\sigma$ representing the NtD map. By the proposed method, L can be chosen to be very small (≤ 3) to yield satisfactory results.

Our task is to find a parameterized mapping $\mathcal{G}_\theta$ to approximate $\mathcal{F}_{L,\mathbb{D}}^{-1}$ (6) by minimizing

$$\mathcal{J}(\theta) := \frac{1}{N}\sum_{k=1}^N \|\mathcal{G}_\theta(\{(g_l^{(k)}, f_l^{(k)})\}_{l=1}^L) - \sigma^{(k)}\|^2, \tag{8}$$

for a suitable norm $\|\cdot\|$. Hyper-parameters N, h, L will affect the finite-dimensional approximation to the infiniteness in the following way: h determines the resolution to approximate D; N affects the representativity of the training data set; L decides how much of a finite portion of the infinite spectral information of Λ_σ can be accessed.

3.2 From harmonic extension to tensor-to-tensor

To establish the connection between the problem of interest with the attention used in the Transformers, we first construct higher-dimensional tensors from the 1D boundary data. The key is a harmonic extension of the boundary data that can be viewed as a PDE-based feature map. We begin with a theorem to motivate it.

Let $\mathcal{I}^D$ be the characteristic function of D named *index function*, i.e., $\mathcal{I}^D(x) = 1$ if $x \in D$ and $\mathcal{I}^D(x) = 0$ if $x \notin D$. Thus, σ can be directly identified by the shape of D through the formula $\sigma = \sigma_1\mathcal{I}^D + \sigma_0(1 - \mathcal{I}^D)$. In this setup, reconstructing σ is equivalent to reconstructing $\mathcal{I}^D$.

Without loss of generality, we let $\sigma_1 > \sigma_0$. Λ_{σ_0} is understood as the NtD map with $\sigma = \sigma_0$ on the whole domain, i.e., it is taken as the known background conductivity (no inclusion), and thus $\Lambda_{\sigma_0} g$ can be readily computed. Then $f - \Lambda_{\sigma_0} g = (\Lambda_\sigma - \Lambda_{\sigma_0})g$ measures the difference between the NtD mappings and encodes the information of σ. The operator $\Lambda_\sigma - \Lambda_{\sigma_0}$ is positive definite, and it has eigenvalues $\{\lambda_l\}_{l=1}^\infty$ with $\lambda_1 > \lambda_2 > \cdots > 0$ (Cheng et al., 1989).

Theorem 1. *Suppose that the 1D boundary data g_l is the eigenfunction of $\Lambda_\sigma - \Lambda_{\sigma_0}$ corresponding to the l-th eigenvalue λ_l, and let the 2D data functions ϕ_l be obtained by solving*

$$-\Delta\phi_l = 0 \quad \text{in} \quad \Omega, \quad \mathbf{n}\cdot\nabla\phi_l = (f_l - \Lambda_{\sigma_0} g_l) \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega}\phi_l \,\mathrm{d}s = 0, \tag{9}$$

for $l = 1, 2, \ldots$. Define the space $\widetilde{\mathbb{S}}_L = \mathrm{Span}\{\partial_{x_1}\phi_l\ \partial_{x_2}\phi_l\ :\ l = 1, \ldots, L\}$, and the dictionary $\mathbb{S}_L = \{a_1 + a_2 \arctan(a_3 v)\ :\ v \in \widetilde{\mathbb{S}}_L,\ a_1, a_2, a_3 \in \mathbb{R}\}$. Then, for any $\epsilon > 0$, we con construct an index function $\mathcal{I}_L^D \in \mathbb{S}_L$ s.t. $\sup_{x\in\Omega} |\mathcal{I}^D(x) - \mathcal{I}_L^D(x)| \leq \epsilon$ provided L is large enough.

The full proof of Theorem 1 can be found in Appendix E. This theorem gives a constructive approach for approximating $\mathcal{F}^{-1}$ and justifies the practice of approximating $\mathcal{I}^D$ when L is large enough. The function ϕ_l is called the *harmonic extension* of $f_l - \Lambda_{\sigma_0} g_l$.

On the other hand, it relies on "knowing" the entire NtD map Λ_σ to construct $\mathcal{I}_L^D$ explicitly. Namely, the coefficients of $\partial_{x_1}\phi_l\ \partial_{x_2}\phi_l$ depend on a big chunk of spectral information (eigenvalues and eigenfunctions) of Λ_σ, which may not be available in practice. Thus, the mathematics itself in this theorem does not provide an architectural hint on building a structure-conforming DNN.

To further dig out the hidden structure, we focus on the case of a single measurement, i.e., $L = 1$. With this setting, it is possible to derive an explicit and simple formula to approximate $\mathcal{I}^D$ which

is achieved by the classical direct sampling methods (DSM) (Chow et al., 2014; 2015; Kang et al., 2018; Ito et al., 2013; Ji et al., 2019; Harris & Kleefeld, 2019; Ahn et al., 2020). For EIT,

$$\mathcal{I}^D(x) \approx \mathcal{I}_1^D(x) := R(x)\left(\mathbf{d}(x)\cdot\nabla\phi(x)\right) \quad x\in\Omega,\ \mathbf{d}(x)\in\mathbb{R}^2, \tag{10}$$

is derived in Chow et al. (2014), where (see a much more detailed formulation in Appendix D)

- ϕ is the harmonic extension of $f-\Lambda_{\sigma_0}g$ with certain noise $f-\Lambda_{\sigma_0}g+\xi$;
- $\mathbf{d}(x)$ is called a probing direction and can be chosen empirically as $\mathbf{d}(x)=\nabla\phi(x)/\|\nabla\phi(x)\|$;
- $R(x) = \left(\|f-\Lambda_{\sigma_0}g\|_{\partial\Omega}|\eta_x|_Y\right)^{-1}$, where η_x is a function of $\mathbf{d}(x)$ and measured in $|\cdot|_Y$ semi-norm on boundary $\partial\Omega$.

Both ϕ and η_x can be computed effectively by traditional fast PDE solvers, such as finite difference or finite element methods based on Ω_h in Section 3.1. However, the reconstruction accuracy is much limited by a single measurement, the nonparametric ansatz, and empirical choices of $\mathbf{d}(x)$ and $|\cdot|_Y$. These restrictions leave room for DL methodology. See Appendix D for a detailed discussion.

Constructing harmonic extension (2D features) from boundary data (1D signal input with limited depth) can contribute to the desired high-quality reconstruction. First, harmonic functions are highly smooth away from the boundary, of which the solution automatically smooths out the noise on the boundary due to PDE theory (Gilbarg & Trudinger, 2001, Chapter 8), and thus make the reconstruction highly robust with respect to the noise (e.g., see Figure 3 in Appendix C.1). Second, in terms of using certain backbone networks to generate features for downstream tasks, harmonic extensions can be understood as a problem-specific way to design higher dimensional feature maps (Álvarez et al., 2012), which renders samples more separable in a higher dimensional data manifold than the one with merely boundary data. See Figure 1 to illustrate this procedure.

The information of σ is deeply hidden in ϕ. As shown in Figure 1 (see also Appendix C), one cannot observe any pattern of σ directly from ϕ. It is different from and more challenging than the inverse problems studied in (Bhattacharya et al., 2021; Khoo et al., 2021) that aim to reconstruct 2D targets from the much more informative 2D internal data of u.

In summary, both Theorem 1 and the formula of DSM (10) offer inspiration to give a potential answer to (Q1): the *harmonic extension* ϕ (2D features) of $f-\Lambda_{\sigma_0}g$ (1D measurements) naturally encodes the information of the true characteristic function $\mathcal{I}_D$ (2D targets). As there is a pointwise correspondence between the harmonic extensions and the targets at 2D grids, a tensor representation of $\nabla\phi_l$ at these grid points can then be used as the input to a *tensor2tensor*-type DNN to learn $\mathcal{I}_D$. Naturally, the grids are set as the positional embedding explicitly. In comparison, the positional information is buried more deeply in 1D measurements. As shown in Figure 1, $\mathcal{G}_\theta$ can be nicely decoupled into a composition of a learnable neural network operator $\mathcal{T}_\theta$ and a non-learnable PDE-based feature map $\mathcal{H}$, i.e., $\mathcal{G}_\theta = \mathcal{T}_\theta\circ\mathcal{H}$. The architecture of $\mathcal{T}_\theta$ shall be our interest henceforth.

3.3 From channels in attention to basis in integral transform

In this subsection, a modified attention mechanism is proposed as the basic block in the tensor2tensor-type mapping introduced in the next two subsections. Its reformulation conforms with one of the most used tools in applied mathematics: the integral transform. In many applications such as inverse problems, the interaction (kernel) does not have any explicit form, which meshes well with DL methodology philosophically. In fact, this is precisely the situation of the EIT problem considered.

Let the input of an encoder attention block be $\mathbf{x}_h\in\mathbb{R}^{M\times c}$ with c channels, then the query Q, key K, value V are generated by three learnable projection matrices $\theta := \{W^Q, W^K, W^V\}\subset\mathbb{R}^{c\times c}$: $\diamond = \mathbf{x}_h W^\diamond$, $\diamond\in\{Q,K,V\}$. Here $c\gg L$ is the number of expanded channels for the latent representations. A modified dot-product attention is proposed as follows:

$$U = \mathrm{Attn}(\mathbf{x}_h) := \alpha\left(\mathrm{nl}_Q(Q)\mathrm{nl}_K(K)^\top\right)V = \alpha\big(\widetilde{Q}\widetilde{K}^\top\big)V\in\mathbb{R}^{M\times c}, \tag{11}$$

where $\mathrm{nl}_Q(\cdot)$ and $\mathrm{nl}_K(\cdot)$ are two learnable normalizations. Different from Nguyen & Salazar (2019); Xiong et al. (2020), this pre-inner-product normalization is applied right before the matrix multiplication of query and key. This practice takes inspiration from the normalization in the index function kernel integral (10) and (20), see also Boyd (2001) where the normalization for orthogonal bases essentially uses the (pseudo)inverse of the Gram matrices. In practice, layer normalization (Ba

et al., 2016) or batch normalization (Ioffe & Szegedy, 2015) is used as a cheap alternative. Constant $\alpha = h^2$ is a mesh-based weight such that the summation becomes an approximation to an integral.

To elaborate these rationales, the j-th column of the i-th row U_i of U is $(U_i)^j = \alpha\, A_{i\bullet} \cdot V^j$, in which the i-th row $A_{i\bullet} = (\widetilde{Q}\widetilde{K}^\top)_{i\bullet}$ and the j-th column $V^j := V_{\bullet j}$. Thus, applying this to every column $1 \le j \le c$, attention (11) becomes a basis expansion representation for the i-th row U_i

$$U_i = \alpha\, A_{i\bullet} \begin{pmatrix} \text{——} & V_1 & \text{——} \\ \text{——} & \vdots & \text{——} \\ \text{——} & V_M & \text{——} \end{pmatrix} = \sum\nolimits_{m=1}^{M} \alpha\, A_{im} V_m =: \sum\nolimits_{m=1}^{M} \mathcal{A}(Q_i, K_m)\, V_m. \tag{12}$$

Here, $\alpha A_{i\bullet}$ contains the coefficients for the linear combination of $\{V_m\}_{m=1}^M$. This set $\{V_m\}_{m=1}^M$ forms the V's row space, and it further forms each row of the output U by multiplying with A. $\mathcal{A}(\cdot,\cdot)$ in (12) stands for the attention kernel, which aggregates the pixel-wise feature maps to measure how the projected latent representations interact. Moreover, the latent representation U_i in an encoder layer is spanned by the row space of V and is being nonlinearly updated cross-layer-wise.

For $\mathbf{x}_h, U, Q, K, V$, a set of feature maps are assumed to exist: for example $u(\cdot)$ maps $\mathbb{R}^2 \to \mathbb{R}^{1\times c}$, i.e., $U_i = u(z_i) = [u_1(z_i), \cdots, u_c(z_i)]$, e.g., see Choromanski et al. (2021), then an instance-dependent kernel $\kappa_\theta(\cdot,\cdot) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ can be defined by

$$\mathcal{A}(Q_i, K_j) := \alpha \langle \widetilde{Q}_i, \widetilde{K}_j \rangle = \alpha \langle q(z_i), k(z_j) \rangle =: \alpha\, \kappa_\theta(z_i, z_j). \tag{13}$$

Now the discrete kernel $\mathcal{A}(\cdot,\cdot)$ with tensorial input is rewritten to this kernel $\kappa_\theta(\cdot,\cdot)$, thus the dot-product attention is expressed as a nonlinear integral transform for the l-th channel:

$$u_l(z) = \alpha \sum\nolimits_{x\in\mathcal{M}} \big(q(z)\cdot k(x)\big) v_l(x)\, \delta_x \approx \int\nolimits_\Omega \kappa_\theta(z,x) v_l(x)\, d\mu(x), \quad 1 \le l \le c. \tag{14}$$

Through certain minimization such as (8), the backpropagation updates θ, which further leads a new set of latent representations. This procedure can be viewed as an iterative method to update the basis residing in each channel by solving the Fredholm integral equation of the first kind in (14).

To connect attention with inverse problems, the multiplicative structure in a kernel integral form for attention (14) is particularly useful. (14) is a type of Pincherle-Goursat (degenerate) kernels (Kress, 1999, Chapter 11) and approximates the full kernel using only a finite number of bases. The number of learned basis functions in expansion (12) depends on the number of channels n. Here we show the following theorem; heuristically, it says that: given enough but finite channels of latent representations, the attention kernel integral can "bootstrap" in the frequency domain, that is, generating an output representation with higher frequencies than the input. Similar approximation results are impossible for layer-wise propagation in CNN if one opts for the usual framelet/wavelet interpretation (Ye et al., 2018). For example, if there are no edge-like local features in the input (see for empirical evidence in Figure 9 and Figure 10), a single layer of CNN filters without nonlinearity cannot learn weights to extract edges. The full proof with a more rigorous setting is in Appendix F.

Theorem 2 (Frequency bootstrapping)**.** *Suppose there exists a channel l in V such that $(V_i)_l = \sin(az_i)$ for some $a \in \mathbb{Z}^+$, the current finite-channel sum kernel $\mathcal{A}(\cdot,\cdot)$ approximates a non-separable kernel to an error of $O(\epsilon)$ under certain norm $\|\cdot\|_X$. Then, there exists a set of weights such that certain channel k' in the output of (12) approximates $\sin(a'z)$, $\mathbb{Z}^+ \ni a' > a$ with an error of $O(\epsilon)$ under the same norm.*

The considered inverse problem is essentially to recover higher-frequency eigenpairs of Λ_σ based on lower-frequency data, see, e.g., Figure 1. Λ_σ together with all its spectral information can be determined by the recovered inclusion shape. Thus, the existence in Theorem 2 partially justifies the advantages of adopting the attention mechanism for the considered problem.

3.4 From index function integral to Transformer

In (10), the probing direction $\mathbf{d}(x)$, the inner product $\mathbf{d}(x)\cdot\nabla\phi(x)$, and the norm $|\cdot|_Y$ are used as ingredients to form certain non-local instance-based learnable kernel integration. This non-localness is a fundamental trait for many inverse problems, in that $\mathcal{I}^D(x)$ depends on the entire data function. Then, the discretization of the modified index function is shown to match the multiplicative structure of the modified attention mechanism in (11).

In the forthcoming derivations, $\mathcal{K}(x,y)$, $\mathcal{Q}(x,y)$, and a self-adjoint positive definite linear operator $\mathcal{V} : L^2(\partial\Omega) \to L^2(\partial\Omega)$, are shown to yield the emblematic Q-K-V structure of attention. To this end, we make the following modifications and assumptions to the original index function in (10).

- The reformulation of the index function is motivated by the heuristics that the agglomerated global information of ϕ could be used as "keys" to locate a point x.
$$\hat{\mathcal{I}}_1^D(x) := R(x)\int_\Omega \mathbf{d}(x)\cdot\mathcal{K}(x,y)\nabla\phi(y)\,\mathrm{d}y. \tag{15}$$
If an ansatz $\mathcal{K}(x,y)=\delta_x(y)$ is adopted, then (15) reverts to the original one in (10).
- The probing direction $\mathbf{d}(x)$ as "query" is reasonably assumed to have a global dependence on ϕ
$$\mathbf{d}(x) := \int_\Omega \mathcal{Q}(x,y)\nabla\phi(y)\,\mathrm{d}y. \tag{16}$$
If $\mathcal{Q}(x,y)=\delta_x(y)/\|\nabla\phi(x)\|$, then $\mathbf{d}(x)=\nabla\phi(x)/\|\nabla\phi(x)\|$ which is the choice of the probing direction in (Ikehata, 2000; Ikehata & Siltanen, 2000; Ikehata, 2007).

- In the quantity $R(x)$ in (10), the key is $|\cdot|_Y$ which is assumed to have the following form:
$$|\eta_x|_Y^2 := (\mathcal{V}\eta_x,\eta_x)_{L^2(\partial\Omega)}. \tag{17}$$
In Chow et al. (2014), it is shown that if $\mathcal{V}$ induces a kernel with sharply peaked Gaussian-like distribution, the index function in (10) can achieve maximum values for points inside D.

Based on the assumptions from (15) to (17), we derive a matrix representation approximating the new index function on a grid, which accords well with an attention-like architecture. Denote by $\boldsymbol{\phi}_n$: the vector that interpolates $\partial_{x_n}\phi$ at the grid points $\{z_j\}$, $n=1,2$.

Here, we sketch the outline of the derivation and present the detailed derivation in Appendix D. We shall discretize the variable x by grid points z_i in (15) and obtain an approximation to the integral:
$$\int_\Omega \mathcal{K}(z_i,y)\partial_{x_n}\phi(y)\,\mathrm{d}y \approx \sum\nolimits_j \omega_j\mathcal{K}(z_i,z_j)\partial_{x_n}\phi(z_j) =: \mathbf{k}_i^T\boldsymbol{\phi}_n, \tag{18}$$
where $\{\omega_j\}$ are some integration quadrature weights. We then consider (16) and focus on one component $d_n(x)$ of $\mathbf{d}(x)$. With a suitable approximated integral, it can be rewritten as
$$d_n(z_i) \approx \sum\nolimits_j \omega_j\mathcal{Q}(z_i,z_j)\partial_{x_n}\phi(z_j) =: \mathbf{q}_i^T\boldsymbol{\phi}_n. \tag{19}$$
Note that the self-adjoint positive definite operator $\mathcal{V}$ in (17) can be parameterized by a symmetric positive definite (SPD) matrix denoted by V. There exist vectors $\mathbf{v}_{n,i}$ such that $|\eta_{z_i}|_Y^2 \approx \sum_n \boldsymbol{\phi}_n^T\mathbf{v}_{n,i}\mathbf{v}_{n,i}^T\boldsymbol{\phi}_n$. Then, the modified indicator function can be written as
$$\hat{\mathcal{I}}_1^D(z_i) \approx \Big\{\|f-\Lambda_{\sigma_0}g\|_{\partial\Omega}^{-1}\big(\sum_n \boldsymbol{\phi}_n^T\mathbf{v}_{n,i}\mathbf{v}_{n,i}^T\boldsymbol{\phi}_n\big)^{-1/2}\Big\}\sum_n \boldsymbol{\phi}_n^T\mathbf{q}_i\mathbf{k}_i^T\boldsymbol{\phi}_n. \tag{20}$$
Now, using the notation from Section 3.3, we denote the learnable kernel matrices and an input vector: for $\diamond\in\{Q,K,V\}$, and $\mathbf{u}\in\{\mathbf{q},\mathbf{k},\mathbf{v}\}$
$$W^\diamond = \begin{bmatrix} \mathbf{u}_{1,1} & \cdots & \mathbf{u}_{1,M} \\ \mathbf{u}_{2,1} & \cdots & \mathbf{u}_{2,M}\end{bmatrix} \in \mathbb{R}^{2M\times M},\quad \mathbf{x}_h = [\ \boldsymbol{\phi}_1 \quad \boldsymbol{\phi}_2\ ] \in \mathbb{R}^{1\times 2M}. \tag{21}$$
Then, we can rewrite (20) as
$$[\hat{\mathcal{I}}_1^D(z_i)]_{i=1}^M \approx C_{f,g}(\mathbf{x}_hW^Q * \mathbf{x}_hW^K)/(\mathbf{x}_hW^V * \mathbf{x}_hW^V)^{1/2} \tag{22}$$
where $C_{f,g}=\|f-\Lambda_{\sigma_0}g\|_{\partial\Omega}^{-1}$ is a normalization weight, and both $*$ and $/$ are element-wise. Here, we may define $Q=\mathbf{x}_hW^Q$, $K=\mathbf{x}_hW^K$, and $V=\mathbf{x}_hW^V$ as the query, keys, and values. We can see that the right matrix multiplications (11) in the attention mechanism are low-rank approximations of the ones above. Hence, based on (22), essentially we need to find a function $\mathbf{I}$ resulting in a vector approximation to the true characteristic function $\{\mathcal{I}^D(z_j)\}$
$$\mathbf{I}(Q,K,V) \approx [\mathcal{I}^D(z_i)]_{i=1}^M. \tag{23}$$
Moreover, when there are L data pairs, the data functions ϕ_l are generated by computing their harmonic extensions as in (9). Then, each ϕ_l is then treated as a channel of the input data $\mathbf{x}_h$.

In summary, the expressions in (22) and (23) reveal that a Transformer may be able to generalize the classical non-parametrized DSM formula further in (10) to non-local learnable kernels. Thus, it may have an intrinsic architectural advantage that handles multiple data pairs. In the subsequent EIT benchmarks, we provide a potential answer to the question (Q2); namely, the attention architecture is better suited for the tasks of reconstruction, as it conforms better with the underlying mathematical structure. The ability to learn global interactions by attention, supported by a non-local kernel interpretation, matches the long-range dependence nature of inverse problems.

4 EXPERIMENTS

In this section we present some experimental results to show the quality of the reconstruction. The benchmark contains sampling of inclusions of random ellipses (targets), and the input data has a ***single*** channel ($L = 1$) of the 2D harmonic extension feature from the 1D boundary measurements. The training uses 1cycle and a mini-batch ADAM for 50 epochs. The evaluated model is taken from the epoch with the best validation metric on a reserved subset. There are several baseline models to compare: the CNN-based U-Nets (Ronneberger et al., 2015; Guo & Jiang, 2020); the state-of-the-art operator learner Fourier Neural Operator (FNO) (Li et al., 2021a) and its variant with a token-mixing layer (Guibas et al., 2022); MultiWavelet Neural Operator (MWO) (Gupta et al., 2021). The Transformer model of interest is a drop-in replacement of the baseline U-Net, and it is named by U-Integral Transformer (UIT). UIT uses the kernel integral inspired attention (11), and we also compare UIT with the linear attention-based Hybrid U-Transformer in Gao et al. (2021), as well as a Hadamard product-based cross-attention U-Transformer in Wang et al. (2022). An ablation study is also conducted by replacing the convolution layers in the U-Net with attention (11) on the coarsest level. For more details of the hyperparameters' setup in the data generation, training, evaluation, network architectures please refer to Section 3.1, Appendix C.1, and Appendix C.2.

The comparison result can be found in Table 1. Because FNO (AFNO, MWO) keeps only the lower modes in spectra, it performs relatively poor in this EIT benchmark where one needs to recover traits that consist of higher modes (sharp boundary edges of inclusion) from lower modes (smooth harmonic extension). Attention-based models are capable to recover "high-frequency target from low-frequency data", and generally outperform the CNN-based U-Nets despite having only $1/3$ of the parameters. Another highlight is that the proposed models are highly robust to noise thanks to the unique PDE-based feature map through harmonic extension. The proposed models can recover the buried domain under a moderately large noise (5%) and an extreme amount of noise (20%) which can be disastrous for many classical methods.

Table 1: Evaluation metrics of the EIT benchmark tests. τ: the normalized relative strength of noises added in the boundary data before the harmonic extension; see Appendix C for details. L^2-error and cross entropy: the closer to 0 the better; Dice coefficient: the closer to 1 the better.

	Relative L^2 error			Position-wise cross entropy			Dice coefficient			# params
	$\tau = 0$	$\tau = 0.05$	$\tau = 0.2$	$\tau = 0$	$\tau = 0.05$	$\tau = 0.2$	$\tau = 0$	$\tau = 0.05$	$\tau = 0.2$	
U-Net baseline	0.200	0.341	0.366	0.0836	0.132	0.143	0.845	0.810	0.799	7.7m
U-Net+Coarse Attn	0.184	0.343	0.360	0.0801	0.136	0.147	0.852	0.807	0.804	8.4m
U-Net big	0.195	0.338	0.350	0.0791	0.133	0.138	0.850	0.812	0.805	31.0m
FNO2d baseline	0.318	0.492	0.502	0.396	0.467	0.508	0.650	0.592	0.582	10.4m
Adaptive FNO2d	0.323	0.497	0.499	0.391	0.466	0.471	0.635	0.595	0.592	10.9m
FNO2d big	0.386	0.482	0.501	0.310	0.465	0.499	0.638	0.601	0.580	33.6m
Multiwavelet NO	0.275	0.390	0.407	0.152	0.178	0.192	0.715	0.694	0.688	9.8m
Hybrid UT	0.185	0.320	0.333	0.0785	0.112	0.116	0.877	0.829	0.821	11.9m
Cross-Attention UT	0.171	0.305	0.311	0.0619	0.105	0.109	0.887	0.840	0.829	11.4m
UIT+Softmax (ours)	**0.159**	**0.261**	**0.269**	**0.0551**	**0.0969**	**0.0977**	**0.903**	**0.862**	**0.848**	11.1m
UIT (ours)	**0.163**	**0.261**	**0.272**	**0.0564**	**0.0967**	**0.0981**	**0.897**	**0.858**	**0.845**	11.4m
UIT+($L=3$) (ours)	**0.147**	**0.250**	**0.254**	**0.0471**	**0.0882**	**0.0900**	**0.914**	**0.891**	**0.880**	11.4m

5 CONCLUSION

For a boundary value inverse problem, we propose a novel operator learner based on the mathematical structure of the inverse operator and Transformer. The proposed architecture consists of two components: the first one is a harmonic extension of boundary data (a PDE-based feature map), and the second one is a modified attention mechanism derived from the classical DSM by introducing learnable non-local integral kernels. The evaluation accuracy on the benchmark problems surpasses the current widely-used CNN-based U-Net and the best operator learner FNO. This research strengthens the insights that the attention is an adaptable neural architecture that can incorporate a priori mathematical knowledge to design more physics-compatible DNN architectures. However, we acknowledge some limitations: in this study, σ to be recovered relies on a piecewise constant assumption. For many EIT applications in medical imaging and industrial monitoring, σ may involve non-sharp transitions or even contain highly anisotropic/multiscale behaviors; see Appendix G for more discussion on limitations and possible approaches.

ACKNOWLEDGMENTS

L. Chen is supported in part by National Science Foundation grants DMS-1913080 and DMS-2012465, and DMS-2132710. S. Cao is supported in part by National Science Foundation grants DMS-1913080 and DMS-2136075. The hardware to perform the experiments are sponsored by NSF grants DMS-2136075, and UMKC School of Science and Engineering computing facilities. No additional revenues are related to this work. The authors would like to thank Ms. Jinrong Wei (University of California Irvine) for the proofreading and various suggestions on the manuscript. The authors would like to thank Dr. Jun Zou (The Chinese University of Hong Kong) and Dr. Bangti Jin (University College London & The Chinese University of Hong Kong) for their comments on inverse problems. The authors also greatly appreciate the valuable suggestions and comments by the anonymous reviewers.

REPRODUCIBILITY STATEMENT

This paper is reproducible. Experimental details about all empirical results described in this paper are provided in Appendix C. Additionally, we provide the PyTorch (Paszke et al., 2019) code for reproducing our results at `https://github.com/scaomath/eit-transformer`. The dataset used in this paper is available at `https://www.kaggle.com/datasets/scaomath/eletrical-impedance-tomography-dataset`. Formal proofs under a rigorous setting of all our theoretical results are provided in Appendices E-F.

REFERENCES

Andy Adler and Robert Guardo. A neural network image reconstruction technique for electrical impedance tomography. *IEEE Trans Med Imaging*, 13(4):594–600, 1994.

Chi Young Ahn, Taeyoung Ha, and Won-Kwang Park. Direct sampling method for identifying magnetic inhomogeneities in limited-aperture inverse scattering problem. *Computers & Mathematics with Applications*, 80(12):2811–2829, 2020. ISSN 0898-1221. URL https://www.sciencedirect.com/science/article/pii/S0898122120304089.

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. *Kernels for Vector-Valued Functions: A Review*. 2012.

Habib Ammari and Hyeonbae Kang. *Reconstruction of Small Inhomogeneities from Boundary Measurements*. Berlin: Springer, 2004.

Habib Ammari and Hyeonbae Kang. *Polarization and Moment Tensors: With Applications to Inverse Problems and Effective Medium Theory*. New York: Springer, 2007.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3): 337–404, 1950.

Kari Astala and Lassi Päivärinta. Calderón's inverse conductivity problem in the plane. *Annals of Mathematics*, 163(1):265–299, 2006a.

Kari Astala and Lassi Päivärinta. Calderón's inverse conductivity problem in the plane. *Ann. of Math.*, pp. 265–299, 2006b.

Mustapha Azzouz, Martin Hanke, Chantal Oesterlein, and Karl Schilcher. The factorization method for electrical impedance tomography data from a new planar device. *International journal of biomedical imaging*, 2007: 83016–83016, 2007. URL https://pubmed.ncbi.nlm.nih.gov/18350126.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Guillaume Bal. Hybrid inverse problems and internal functionals. *Inverse problems and applications: inside out. II*, 60:325–368, 2013.

Gang Bao, Xiaojing Ye, Yaohua Zang, and Haomin Zhou. Numerical solution of inverse problems by weak adversarial networks. *Inverse Probl.*, 36(11):115003, nov 2020. URL https://doi.org/10.1088/1361-6420/abb447.

George Barbastathis, Aydogan Ozcan, and Guohai Situ. On the use of deep learning for computational imaging. *Optica*, 6(8):921–943, Aug 2019. URL http://www.osapublishing.org/optica/abstract.cfm?URI=optica-6-8-921.

Bartolomé Barceló, Eugene Fabes, and Jin K. Seo. The inverse conductivity problem with one measurement: Uniqueness for convex polyhedra. *Proc. Amer. Math. Soc.*, 122(1):183–189, 1994.

Hanene Ben Yedder, Aïcha BenTaieb, Majid Shokoufi, Amir Zahiremami, Farid Golnaraghi, and Ghassan Hamarneh. Deep learning based image reconstruction for diffuse optical tomography. In Florian Knoll, Andreas Maier, and Daniel Rueckert (eds.), *Machine Learning for Medical Image Reconstruction*, pp. 112–119, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00129-2.

Jose Antonio Lara Benitez, Takashi Furuya, Florian Faucher, Xavier Tricoche, and Maarten V de Hoop. Fine-tuning neural-operator architectures for training and generalization. *arXiv preprint arXiv:2301.11509*, 2023.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model reduction and neural networks for parametric pdes. *The SMAI journal of computational mathematics*, 7, 2021.

Oussama Boussif, Yoshua Bengio, Loubna Benabbou, and Dan Assouline. MAgnet: Mesh agnostic neural PDE solver. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=bx2roi8hca8.

John P Boyd. *Chebyshev and Fourier spectral methods*. Courier Corporation, 2001.

Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K Gupta. Clifford neural layers for PDE modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=okwxL_c4x84.

Martin Brühl. Explicit characterization of inclusions in electrical impedance tomography. *SIAM Journal on Mathematical Analysis*, 32(6):1327–1341, 2001. URL https://doi.org/10.1137/S003614100036656X.

Tan Bui-Thanh, Qin Li, and Leonardo Zepeda-Núñez. Bridging and improving theoretical and computational electrical impedance tomography via data completion. *SIAM Journal on Scientific Computing*, 44(3):B668–B693, 2022.

Alberto P. Calderón. On an inverse boundary value problem. *Comput. Appl. Math*, 25(2-3), 2006.

Shuhao Cao. Choose a Transformer: Fourier or Galerkin. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*, volume 34, pp. 24924–24940, 2021. URL https://openreview.net/forum?id=ssohLcmn4-r.

Tony F. Chan and Xue-Cheng Tai. Identification of discontinuous coefficients in elliptic problems using total variation regularization. *SIAM J. Sci. Comput*, 25(3):881–904, 2003.

Dongdong Chen, Julián Tachella, and Mike E Davies. Equivariant imaging: Learning beyond the range space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4379–4388, 2021.

Junqing Chen, Ying Liang, and Jun Zou. Mathematical and numerical study of a three-dimensional inverse eddy current problem. *SIAM J. on Appl. Math.*, 80(3):1467–1492, 2020.

Margaret Cheney. The linear sampling method and the MUSIC algorithm. *Inverse Probl.*, 17(4):591–595, jul 2001. URL https://doi.org/10.1088/0266-5611/17/4/301.

Kuo-Sheng Cheng, David Isaacson, JC Newell, and David G Gisser. Electrode models for electric current computed tomography. *IEEE. Trans. Biomed. Eng.*, 36(9):918–924, 1989.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with Performers. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Yat Tin Chow, Kazufumi Ito, and Jun Zou. A direct sampling method for electrical impedance tomography. *Inverse Probl.*, 30(9):095003, 2014.

Yat Tin Chow, Kazufumi Ito, Keji Liu, and Jun Zou. Direct sampling method for diffusive optical tomography. *SIAM J. Sci. Comput.*, 37(4):A1658–A1684, 2015.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Adam Coxson, Ivo Mihov, Ziwei Wang, Vasil Avramov, Frederik Brooke Barnes, Sergey Slizovskiy, Ciaran Mullan, Ivan Timokhin, David Sanderson, Andrey Kretinin, et al. Machine learning enhanced electrical impedance tomography for 2d materials. *Inverse Problems*, 38(8):085007, 2022.

J. P. Culver, R. Choe, M. J. Holboke, L. Zubkov, T. Durduran, A. Slemp, V. Ntziachristos, B. Chance, and A. G. Yodh. Three-dimensional diffuse optical tomography in the parallel plane transmission geometry: Evaluation of a hybrid frequency domain/continuous wave clinical system for breast imaging. *Medical Physics*, 30(2): 235–247, 2003. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1534109.

Maarten de Hoop, Daniel Zhengyu Huang, Elizabeth Qian, and Andrew M Stuart. The cost-accuracy trade-off in operator learning with neural networks. *Journal of Machine Learning*, 1(3):299–341, June 2022a. URL https://doi.org/10.4208/jml.220509.

Maarten V de Hoop, Matti Lassas, and Christopher A Wong. Deep learning architectures for nonlinear operator functions and nonlinear inverse problems. *Mathematical Statistics and Learning*, 4(1):1–86, February 2022b. URL https://doi.org/10.4171/msl/28.

David C Dobson and Fadil Santosa. An image-enhancement technique for electrical impedance tomography. *Inverse Probl.*, 10(2):317, 1994.

Yuwei Fan and Lexing Ying. Solving electrical impedance tomography with deep learning. *J. Comput. Phys.*, 404:109119, 2020.

Yuwei Fan, Cindy Orozco Bohorquez, and Lexing Ying. BCR-Net: A neural network based on the nonstandard wavelet form. *J. Comput. Phys.*, 384:1–15, 2019. ISSN 0021-9991. URL https://www.sciencedirect.com/science/article/pii/S0021999119300762.

Jinchao Feng, Qiuwan Sun, Zhe Li, Zhonghua Sun, and Kebin Jia. Back-propagation neural network-based reconstruction algorithm for diffuse optical tomography. *Journal of Biomedical Optics*, 24(5):1 – 12, 2018. URL https://doi.org/10.1117/1.JBO.24.5.051407.

Xosé Fernández-Fuentes, David Mera, Andrés Gómez, and Ignacio Vidal-Franco. Towards a fast and accurate eit inverse problem solver: A machine learning approach. *Electronics*, 7(12), 2018. ISSN 2079-9292. URL https://www.mdpi.com/2079-9292/7/12/422.

Antonio H de O Fonseca, Emanuele Zappala, Josue Ortega Caro, and David van Dijk. Continuous spatiotemporal transformers. *arXiv preprint arXiv:2301.13338*, 2023.

Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 61–71. Springer, 2021.

Omar Ghattas and Karen Willcox. Learning physics-based models from data: perspectives from inverse problems and model reduction. *Acta Numerica*, 30:445–554, 2021.

David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, New York, 2 edition, 2001.

H Griffiths, WR Stewart, and W Gough. Magnetic induction tomography. a measuring system for biological tissues. *Ann N Y Acad Sci.*, 20(873), 1999.

Danping Gu, Dong Liu, Danny Smyl, Jiansong Deng, and Jiangfeng Du. Supershape recovery from electrical impedance tomography data. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive Fourier neural operators: Efficient token mixers for Transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EXHG-A3jlM.

Ruchi Guo and Jiahua Jiang. Construct deep neural networks based on direct sampling methods for solving electrical impedance tomography. *SIAM J. Sci. Comput.*, 43(3):B678–B711, 2020.

Ruchi Guo, Tao Lin, and Yanping Lin. A fixed mesh method with immersed finite elements for solving interface inverse problems. *J. Sci. Comput.*, 79(1):148–175, 2019.

Ruchi Guo, Jiahua Jiang, and Yi Li. Learn an index operator by CNN for solving diffusive optical tomography: a deep direct sampling method. *J. Sci. Comput.*, 2021.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. In *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021. URL https://openreview.net/forum?id=LZDiWaC9CGL.

Gaurav Gupta, Xiongye Xiao, Radu Balan, and Paul Bogdan. Non-linear operator approximations for initial value problems. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=d2TT6gK9qZn.

Sarah J Hamilton, Asko Hänninen, Andreas Hauptmann, and Ville Kolehmainen. Beltrami-net: domain-independent deep d-bar learning for absolute imaging with electrical impedance tomography (a-eit). *Physiol Meas.*, 40(7):074002, 2019.

Sarah Jane Hamilton and Andreas Hauptmann. Deep d-bar: Real-time electrical impedance tomography imaging with deep neural networks. *IEEE Trans Med Imaging*, 37(10):2367–2377, 2018.

Xing Han, Tongzheng Ren, Tan Minh Nguyen, Khai Nguyen, Joydeep Ghosh, and Nhat Ho. Robustify transformers with robust kernel density estimation. *arXiv preprint arXiv:2210.05794*, 2022.

Martin Hanke and Martin Brühl. Recent progress in electrical impedance tomography. 19(6):S65–S90, nov 2003. URL https://doi.org/10.1088/0266-5611/19/6/055.

Isaac Harris and Andreas Kleefeld. Analysis of new direct sampling indicators for far-field measurements. *Inverse Problems*, 35(5):054002, apr 2019. URL https://doi.org/10.1088/1361-6420/ab08be.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, Los Alamitos, CA, USA, jan 2022. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/WACV51458.2022.00181.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sheng He, P. Ellen Grant, and Yangming Ou. Global-local transformer for brain age estimation. *IEEE Transactions on Medical Imaging*, 41(1):213–224, 2022.

David S Holder. *Electrical impedance tomography: methods, history and applications*. CRC Press, 2004.

Radek Hrabuska, Michal Prauzek, Marketa Venclikova, and Jaromir Konecny. Image reconstruction for electrical impedance tomography: Experimental comparison of radial basis neural network and gauss – newton method. *IFAC-PapersOnLine*, 51(6):438–443, 2018. ISSN 2405-8963. URL https://www.sciencedirect.com/science/article/pii/S2405896318308589. 15th IFAC Conference on Programmable Devices and Embedded Systems PDeS 2018.

Peiyan Hu, Qi Meng, Bingguang Chen, Shiqi Gong, Yue Wang, Wei Chen, Rongchan Zhu, Zhi-Ming Ma, and Tie-Yan Liu. Neural operator with regularity structure for modeling dynamics driven by spdes. *arXiv preprint arXiv:2204.06255*, 2022.

Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, pp. 4533–4543. PMLR, 2021.

Masaru Ikehata. Reconstruction of the support function for inclusion from boundary measurements. *Journal of Inverse and Ill-posed Problems*, 8(4):367–378, 2000. URL https://doi.org/10.1515/jiip.2000.8.4.367.

Masaru Ikehata. Probe method and a carleman function. *Inverse Problems*, 23(5):1871–1894, aug 2007. URL https://doi.org/10.1088/0266-5611/23/5/006.

Masaru Ikehata and Samuli Siltanen. Numerical method for finding the convex hull of an inclusion in conductivity from boundary measurements. *Inverse Problems*, 16(4):1043–1052, aug 2000. URL https://doi.org/10.1088/0266-5611/16/4/311.

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

Haris Iqbal. Harisiqbal88/plotneuralnet v1.0.0, 2018. URL https://zenodo.org/record/2526396.

Victor Isakov and Jeffrey Powell. On the inverse conductivity problem with one measurement. *Inverse Probl.*, 6: 311, 1990.

Kazufumi Ito, Bangti Jin, and Jun Zou. A direct sampling method for inverse electromagnetic medium scattering. *Inverse Probl.*, 29(9):095018, sep 2013. URL https://doi.org/10.1088/0266-5611/29/9/095018.

Xia Ji, Xiaodong Liu, and Bo Zhang. Phaseless inverse source scattering problem: Phase retrieval, uniqueness and direct sampling methods. *Journal of Computational Physics: X*, 1:100003, 2019. ISSN 2590-0552. URL https://www.sciencedirect.com/science/article/pii/S2590055219300022.

Bangti Jin and Yifeng Xu. Adaptive reconstruction for electrical impedance tomography with a piecewise constant conductivity. *Inverse Problems*, 36(1):014003, 2019.

Bangti Jin, Xiyao Li, and Xiliang Lu. Imaging conductivity from current density magnitude using neural networks. *Inverse Problems*, 38(7):075003, 2022a.

Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

Pengzhan Jin, Shuai Meng, and Lu Lu. Mionet: Learning multiple-input operators via tensor product. *arXiv preprint arXiv:2202.06137*, 2022b.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. 2016.

Eunhee Kang, Junhong Min, and Jong Chul Ye. A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. *Med Phys*, 44(10):e360–e375, Oct 2017. ISSN 2473-4209 (Electronic); 0094-2405 (Linking).

Hyeonbae Kang and Jin Keun Seo. Note on uniqueness and stability for the inverse conductivity problems with one measurement. *J. Korean Math. Soc*, pp. 781–792, 2001.

Sangwoo Kang, Marc Lambert, and Won-Kwang Park. Direct sampling method for imaging small dielectric inhomogeneities: analysis and improvement. *Inverse Probl.*, 34(9):095005, jul 2018. URL https://doi.org/10.1088/1361-6420/aacf1d.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

Yuehaw Khoo, Jianfeng Lu, and Lexing Ying. Solving parametric pde problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

Andreas Kirsch and Natalia Grinberg. *The factorization method for inverse problems*, volume 36. OUP Oxford, 2007.

Georgios Kissas, Jacob Seidman, Leonardo Ferreira Guilhoto, Victor M Preciado, George J Pappas, and Paris Perdikaris. Learning operators with coupled attention. *arXiv preprint arXiv:2201.01032*, 2022.

Grzegorz Kłosowski and Tomasz Rymarczyk. Using neural networks and deep learning algorithms in electrical impedance tomography. *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, 7(3), 2017.

Kim Knudsen, Matti Lassas, Jennifer L. Mueller, and Samuli Siltanen. D-bar method for electrical impedance tomography with discontinuous conductivities. *SIAM Journal on Applied Mathematics*, 67(3):893–913, 2007. URL https://doi.org/10.1137/060656930.

Kim Knudsen, Matti Lassas, Jennifer L. Mueller, and Samuli Siltanen. Regularized d-bar method for the inverse conductivity problem. *Inverse Probl. & Imaging*, 3(4):599–624, 2009.

Robert Kohn and Michael Vogelius. Determining conductivity by boundary measurements. *Communications on Pure and Applied Mathematics*, 37(3):289–298, 1984.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for Fourier neural operators. *The Journal of Machine Learning Research*, 22(1):13237–13312, 2021.

Rainer Kress. *Linear Integral Equations*. Springer New York, 1999. URL https://doi.org/10.1007/978-1-4612-0559-3.

Jahanzaib Latif, Chuangbai Xiao, Azhar Imran, and Shanshan Tu. Medical imaging using machine learning and deep learning algorithms: A review. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1–5, 2019.

Okkyun Lee, Jongmin Kim, Yoram Bresler, and Jong Chul Ye. Diffuse optical tomography using generalized music algorithm. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1142–1145, 2011.

Seungjun Lee. Mesh-independent operator learning for PDEs using set representations, 2023. URL https://openreview.net/forum?id=7d-d0BFz6Hf.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *arXiv preprint arXiv:2205.13671*, 2022a.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *arXiv preprint arXiv:2111.03794*, 2021b.

Zongyi Li, Miguel Liu-Schiaffini, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Learning chaotic dynamics in dissipative systems. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=1C36tFZn7sR.

William RB Lionheart. Eit reconstruction algorithms: pitfalls, challenges and recent developments. *Physiological measurement*, 25(1):125, 2004.

Xinliang Liu, Bo Xu, and Lei Zhang. Ht-net: Hierarchical transformer based operator learning model for multiscale pdes. *arXiv preprint arXiv:2210.10890*, 2022.

Fei Lu, Qingci An, and Yue Yu. Nonparametric learning of kernels in nonlocal operators. *arXiv preprint arXiv:2205.11006*, 2022.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George E. Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3): 218–229, 2021. URL https://doi.org/10.1038/s42256-021-00302-5.

Björn Lütjens, Catherine H Crawford, Campbell D Watson, Christopher Hill, and Dava Newman. Multiscale neural operator: Learning fast and grid-independent pde solvers. *arXiv preprint arXiv:2207.11417*, 2022.

Jose Marroquin, Sanjoy Mitter, and Tomaso Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the american statistical association*, 82(397):76–89, 1987.

Thierry Martin and Jérôme Idier. A FEM-based nonlinear map estimator in electrical impedance tomography. In *Proceedings of ICIP*, volume 2, pp. 684–687. IEEE, 1997.

James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

Marketa Michalikova, Rawia Abed, Michal Prauzek, and Jiri Koziorek. Image reconstruction in electrical impedance tomography using neural network. In *2014 Cairo International Biomedical Engineering Conference (CIBEC)*, pp. 39–42. IEEE, 2014.

Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, pp. 154–168. Springer, 2006.

Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

Adrian Nachman, Alexandru Tamasan, and Alexandre Timonov. Conductivity imaging with a single measurement of boundary and interior data. *Inverse Problems*, 23(6):2551, 2007.

Adrian I. Nachman. Global uniqueness for a two-dimensional inverse boundary value problem. *Annals of Mathematics*, 143(1):71–96, 1996.

Tan M. Nguyen, Vai Suliafu, Stanley J. Osher, Long Chen, and Bao Wang. FMMformer: Efficient and Flexible Transformer via Decomposed Near-field and Far-field Attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tan Minh Nguyen, Minh Pham, Tam Minh Nguyen, Khai Nguyen, Stanley Osher, and Nhat Ho. Fourierformer: Transformer meets generalized fourier integral theorem. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=PRd7VG_ki_.

Thanh C. Nguyen, Vy Bui, and George Nehmetallah. Computational optical tomography using 3-D deep convolutional neural networks. *Optical Engineering*, 57(4):1 – 11, 2018. URL https://doi.org/10.1117/1.OE.57.4.043111.

Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, November 2-3 2019. Association for Computational Linguistics. URL https://aclanthology.org/2019.iwslt-1.17.

Yong Zheng Ong, Zuowei Shen, and Haizhao Yang. Integral autoencoder network for discretization-invariant learning. *Journal of Machine Learning Research*, 23(286):1–45, 2022.

Samira Pakravan, Pouria A Mistani, Miguel A Aragon-Calvo, and Frederic Gibou. Solving inverse-pde problems with physics-aware neural networks. *Journal of Computational Physics*, 440:110414, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 8024–8035. 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QtTKTdVrFBB.

Shehan Perera, Srikar Adhikari, and Alper Yilmaz. Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound. In *ICIP*, 2021.

Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, pp. 267–276, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-87588-6. URL https://doi.org/10.1007/978-3-030-87589-3_28.

M. K. Pidcock, M. Kuzuoglu, and K. Leblebicioglu. Analytic and semi-analytic solutions in electrical impedance tomography. i. two-dimensional problems. *Physiological Measurement*, 16(2):77–90, may 1995. URL https://doi.org/10.1088%2F0967-3334%2F16%2F2%2F001.

Shangjie Ren, Kai Sun, Chao Tan, and Feng Dong. A two-stage deep learning method for robust shape reconstruction with electrical impedance tomography. *IEEE Transactions on Instrumentation and Measurement*, 69(7):4887–4897, 2020.

Luca Rondi and Fadil Santosa. Enhanced electrical impedance tomography via the mumford-shah functional. *ESAIM: COCV*, 6:517–538, 2001. URL https://doi.org/10.1051/cocv:2001121.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Tim De Ryck and Siddhartha Mishra. Generic bounds on the approximation error for physics-informed (and) operator learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=bF4eYy3LTR9.

Jacob H Seidman, Georgios Kissas, Paris Perdikaris, and George J. Pappas. NOMAD: Nonlinear manifold decoders for operator learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=5OWV-sZvMl.

Jin Keun Seo, Kang Cheol Kim, Ariungerel Jargal, Kyounghun Lee, and Bastian Harrach. A learning-based method for solving ill-posed nonlinear inverse problems: a simulation study of lung eit. *SIAM journal on Imaging Sciences*, 12(3):1275–1295, 2019.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pp. 1100612. International Society for Optics and Photonics, 2019.

Diping Song, Bin Fu, Fei Li, Jian Xiong, Junjun He, Xiulan Zhang, and Yu Qiao. Deep relation transformer for diagnosing glaucoma with optical coherence tomography and visual field function. *IEEE Transactions on Medical Imaging*, 40(9):2392–2402, 2021.

John Sylvester and Gunther Uhlmann. A global uniqueness theorem for an inverse boundary value problem. *Annals of Mathematics*, 125(1):153–169, 1987.

Chao Tan, Shuhua Lv, Feng Dong, and Masahiro Takei. Image reconstruction based on convolutional neural network for electrical resistance tomography. *IEEE Sensors Journal*, 19(1):196–204, 2018.

Leonardo Tanzi, Andrea Audisio, Giansalvo Cirrincione, Alessandro Aprato, and Enrico Vezzetti. Vision transformer for femur fracture classification. *Injury*, 2022. ISSN 0020-1383. URL https://www.sciencedirect.com/science/article/pii/S0020138322002868.

T. Tarvainen, M. Vauhkonen, and S.R. Arridge. Gauss–newton reconstruction method for optical tomography using the finite element solution of the radiative transfer equation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 109(17):2767–2778, 2008. ISSN 0022-4073. URL https://www.sciencedirect.com/science/article/pii/S0022407308001854.

J Nasehi Tehrani, Alistair McEwan, Craig Jin, and A Van Schaik. L1 regularization method in electrical impedance tomography by using the l1-curve (pareto frontier curve). *Applied Mathematical Modelling*, 36 (3):1095–1105, 2012. ISSN 0307-904X. URL https://www.sciencedirect.com/science/article/pii/S0307904X11004537.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4344–4353, Hong Kong, China, November 2019. URL https://www.aclweb.org/anthology/D19-1443.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30, 2017.

Päivi J Vauhkonen, Marko Vauhkonen, Tuomo Savolainen, and Jari P Kaipio. Three-dimensional electrical impedance tomography based on the complete electrode model. *IEEE Trans. Biomedical Engrg.*, 46(9): 1150–1160, 1999.

Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2390–2394. IEEE, 2022.

Huan Wang, Kai Liu, Yang Wu, Song Wang, Zheng Zhang, Fang Li, and Jiafeng Yao. Image reconstruction for electrical impedance tomography using radial basis function neural network based on hybrid particle swarm optimization algorithm. *IEEE Sensors Journal*, 21(2):1926–1934, 2021a.

Qi Wang, Huaxiang Wang, Ronghua Zhang, Jinhai Wang, Yu Zheng, Ziqiang Cui, and Chengyi Yang. Image reconstruction based on l1 regularization and projection methods for electrical impedance tomography. *Review of Scientific Instruments*, 83(10):104707, 2012. URL https://doi.org/10.1063/1.4760253.

Sifan Wang, Hanwen Wang, and Paris Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed deeponets. *Science advances*, 7(40):eabi8605, 2021b.

Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. U-fno—an enhanced fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163: 104180, 2022.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pp. 10524–10533. PMLR, 2020.

Jong Chul Ye, Yoseob Han, and Eunju Cha. Deep convolutional framelets: A general deep learning framework for inverse problems. *SIAM Journal on Imaging Sciences*, 11(2):991–1048, 2018.

Rui Zhang, Qi Meng, Rongchan Zhu, Yue Wang, Wenlei Shi, Shihua Zhang, Zhi-Ming Ma, and Tie-Yan Liu. Monte carlo neural operator for learning pdes via probabilistic representation. *arXiv preprint arXiv:2302.05104*, 2023.

Jiawei Zhao, Robert Joseph George, Yifei Zhang, Zongyi Li, and Anima Anandkumar. Incremental fourier neural operator. *arXiv preprint arXiv:2211.15188*, 2022.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *arXiv:2109.03201v6*, 2021.

Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, Mar 2018. ISSN 1476-4687 (Electronic); 0028-0836 (Linking).

Jianqing Zhu, Juncai He, and Qiumei Huang. An enhanced v-cycle mgnet model for operator learning in numerical partial differential equations. *arXiv preprint arXiv:2302.00938*, 2023a.

Jianqing Zhu, Juncai He, Lian Zhang, and Jinchao Xu. Fv-mgnet: Fully connected v-cycle mgnet for interpretable time series forecasting. *arXiv preprint arXiv:2302.00962*, 2023b.

A TABLE OF NOTATIONS

Table 2: Notations used in an approximate chronological order and their meaning in this work.

Notation	Meaning
Ω	an underlying spacial domain in $\mathbb{R}^2$
D	a subdomain in Ω (not necessarily topologically-connected)
∂D, $\partial\Omega$	D's and Ω's boundary, 1-dimensional manifolds
∇u	the gradient vector of a function, $\nabla u(x) = (\partial_{x_1} u(x), \partial_{x_2} u(x))$
$\|\cdot\|_\omega$	the L^2-norm on a region ω
$\|\cdot\| = \|\cdot\|_\Omega$	the L^2-norm on whole domain Ω
δ_x	the delta function such that $\int_\Omega f(y)\delta_x(y)\,\mathrm{d}y = f(x), \forall f$.
$\mathbf{n}$	the unit outer normal vector on the boundary $\partial\Omega$.
$\nabla u \cdot \mathbf{n}$	normal derivative of u, measures the rate of change along the direction of $\mathbf{n}$
Λ_σ	NtD map from Neumann data $g := \nabla u \cdot \mathbf{n}$ (how fast the solution changes toward the outward normal direction) to Dirichlet data $f := u\|_{\partial\Omega}$ (the solution's value along the tangential direction)
$H^s(\Omega)$, $s \geq 0$	the Sobolev space of functions
$H^s(\Omega)$, $s < 0$	the bounded linear functional defined on $H^s(\Omega)$
$H_0^s(\Omega)$	all $u \in H^s(\Omega)$ such that u's integral on Ω vanishes
$\|\cdot\|_Y$	the seminorm defined for functions in Y

B BACKGROUND OF EIT

For EIT, an immediate question is whether $\mathcal{F}^{-1}$ and $\mathcal{F}_L^{-1}$ in (4) and (5) are well-defined, namely whether σ can be uniquely determined. In fact, for the case of full measurements ($L = \infty$), the uniqueness for $\mathcal{F}^{-1}$ has been well established, (Brühl, 2001; Hanke & Brühl, 2003; Astala & Päivärinta, 2006a;b; Nachman, 1996; Kohn & Vogelius, 1984; Sylvester & Uhlmann, 1987). It is worthwhile to point out that, in this case, σ is not necessarily a piecewise constant function. In (1), we present a simplified case for purposes of illustrating as well as benchmarking. In general, with the full spectral information of the NtD map, σ can be uniquely determined as a general positive function.

If infinitely many eigenpairs are known, then the operator itself can be precisely characterized using infinitely many feature channels by Reproducing Kernel Hilbert Space (RKHS) theory, e.g., Mercer (1909); Aronszajn (1950); Minh et al. (2006); Morris (2015); Kadri et al. (2016); Lu et al. (2022). In the context of EIT, this is known as the "full measurement". A more challenging and practical problem is to recover σ from only finitely many boundary data pairs. A common practice for the theoretical study of reconstruction using finite measurements is the assumption of σ being a piecewise constant function. The task is usually set to recover the shape and location of the inclusion D. Otherwise, the problem is too ill-posed. With finite measurements, the uniqueness of the inclusion remains a long-standing theoretical open problem, and it can be only established for several special classes of the inclusion shape, such as the convex cylinders in Isakov & Powell (1990) or convex polyhedrons in Barceló et al. (1994). We refer readers to some counter-examples in Kang & Seo (2001) where a two- or three-dimensional ball may not be identified uniquely by one single measurement if the values of σ_0 and σ_1 are unknown.

Furthermore, here we provide one example to illustrate the difficulty in the reconstruction procedure (Pidcock et al., 1995). Let Ω be a unit circle, let D be a circle with the radius $\rho < 1$, and define

$$\sigma(x) = \begin{cases} 1 & \text{if } \|x\| \geq \rho, \\ \sigma_1 & \text{if } \|x\| < \rho, \end{cases} \tag{24}$$

with $\sigma_1 < 1$ being an arbitrary constant. In this case, the eigenpairs of Λ_σ can be explicitly calculated

$$\lambda_l = \frac{1}{l}\frac{1-\rho^{2l}\mu}{1+\rho^{2l}\mu}, \quad \nu_m = \frac{1}{\sqrt{2\pi}}\cos(l\theta) \ \text{ or } \ \frac{1}{\sqrt{2\pi}}\sin(l\theta), \ \ l = 1, 2, ..., \tag{25}$$

with $\mu = (1-\sqrt{\sigma_1})/(1+\sqrt{\sigma_1})$, which are exactly the Fourier modes in a unit circle. In this case, if the set of basis $\{g_l\}_{l=1}^{\infty}$ is just chosen as $\{\cos(\theta), \sin(\theta), \cos(2\theta), \sin(2\theta), ...\}$, the matrix representation $\mathbf{A}_\sigma$ in (3) can be written as an infinite diagonal matrix

$$\mathbf{A}_\sigma \begin{bmatrix} (1-\rho^2\mu)/(1+\rho^2\mu) & 0 & 0 & 0 \\ 0 & (1-\rho^4\mu)/(1+\rho^4\mu) & 0 & 0 \\ 0 & 0 & (1-\rho^4\mu)/(1+\rho^4\mu) & 0 \\ 0 & 0 & 0 & \ddots \end{bmatrix}. \tag{26}$$

Thanks to this special geometry, eigenvalues in (25) can clearly determine ρ and σ_1 as follows:

$$\rho = \sqrt{\frac{(1-\lambda_2)(1+\lambda_1)}{(1+\lambda_2)(1-\lambda_1)}} \quad \text{and} \quad \mu = \frac{(1-\lambda_1)^2(1+\lambda_2)}{(1+\lambda_1)^2(1-\lambda_2)}. \tag{27}$$

However, in practice, $\mathbf{A}_\sigma$ does not have such a simple structure. Approximating $\mathbf{A}_\sigma$ itself requires a large number of data pairs that are not available in the considered case. Besides, an accurate approximation of the eigenvalues of $\mathbf{A}_\sigma$ is also very expensive. Furthermore, for complex inclusion shapes, two eigenvalues are not sufficient to exactly recover the shape and conductivity values.

C Experiment Set-up

C.1 Data generation and training

In the numerical examples, the data generation mainly follows standard practice in theoretical prototyping for solving EIT problems, see e.g.,Chow et al. (2014); Michalikova et al. (2014); Hamilton & Hauptmann (2018); Guo & Jiang (2020); Fan & Ying (2020). For examples, please refer to Figure 2. The computational domain is set to be $\Omega := (-1,1)^2$, and the two media with the different conductivities are with $\sigma_1 = 10$ (inclusion) and $\sigma_0 = 1$ (background). The inclusions are four random ellipses. The lengths of the semi-major axis and semi-minor axis of these ellipses are sampled from $\mathcal{U}(0.1, 0.2)$ and $U(0.2, 0.4)$, respectively. The rotation angles are sampled from $\mathcal{U}(0, 2\pi)$. There are 10800 samples in the training set, from which 20% are reserved as validation. There are 2000 in the testing set for evaluation.

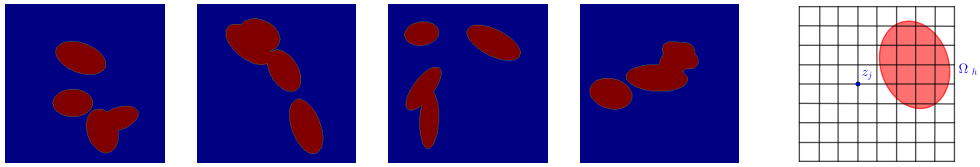

Figure 2: Randomly selected samples of elliptic inclusion to represent the coefficient σ (left 1-4). A Cartesian mesh Ω_h with a grid point $\{z_j\}$ (right). In computation, discretization of $\mathcal{I}^D$ consists of values taken as 1 at the mesh points of Ω_h inside D and 0 at others.

The noise $\xi = \xi(x)$ below (10) is assumed to be

$$\xi(x) = (f(x) - \Lambda_{\sigma_0} g(x))\tau G(x) \tag{28}$$

where τ specifies the relative strength of noise, and $G(x)$ is a normal Gaussian distribution independent with respect to x. As $\xi(x)$ is merely pointwise imposed, the boundary data can be highly rough, even if the ground truth $f(\cdot) - \Lambda_{\sigma_0} g(\cdot)$ is chosen to be smooth. Nevertheless, the harmonic extension makes the noise from boundary data have a minimal impact on the overall reconstruction, thanks to the smoothing property of the inverse of the Laplacian operator $(-\Delta)^{-1}$; for example, please refer to Figure 3. In data generation, the harmonic extension feature map is approximated by finite element methods incorporating stencil modification near the inclusion interfaces (Guo et al., 2019). Similar data augmentation practices using internal data can be found in Nachman et al. (2007); Bal (2013); Jin et al. (2022a).

Thanks to the position-wise binary nature of $\mathcal{I}^D$, another choice of the loss function during training can be the binary cross entropy $\mathcal{L}(\cdot,\cdot)$, applied for a function in $\mathbb{P}$, to measure the distance between the ground truth and the network's prediction

$$\mathcal{L}(\mathbf{p}_h, \mathbf{u}_h) := -\sum\nolimits_{z\in\mathcal{M}} \big(\mathbf{p}_h(z)\ln(\mathbf{u}_h(z)) + (1-\mathbf{p}_h(z))\ln(1-\mathbf{u}_h(z))\big). \tag{29}$$

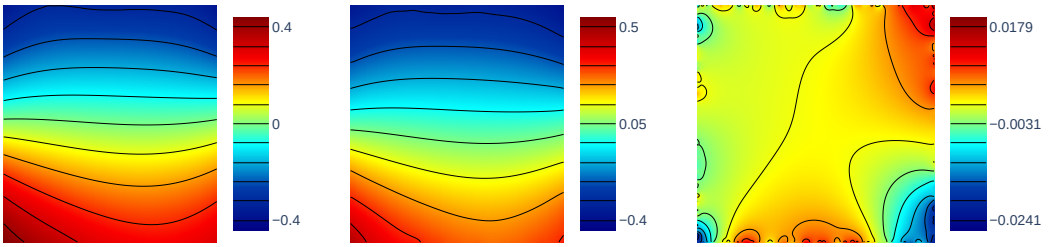

Figure 3: The harmonic extensions ϕ_l with zero noise on boundary (left); the harmonic extensions $\tilde{\phi}_l$ with 20% Gaussian noise on boundary (middle). Their pointwise difference (right). $\|\phi_l - \tilde{\phi}_l\|/\|\phi_l\| = 0.0203$.

Thanks to the Pinsker inequality (e.g., see (Cover, 1999, Section 11.6)), $\mathcal{L}(\mathbf{p}_h, \mathbf{u}_h)$ serves as a good upper bound for the square of the total variation, which can be further bounded below by the L^2-error given the boundedness of the position-wise value.

The training uses `1cycle` (Smith & Topin, 2019) learning rate strategy with a warm-up phase. A mini-batch ADAM iterations are run for 50 epochs with no extra regularization, such as weight decay. The evaluated model is taken from the epoch that has the best validation metric. The learning rate starts and ends with $10^{-3} \cdot lr_{\max}$, and reaches the maximum of $lr_{\max}$ at the end of the 10-th epoch. The $lr_{\max} = 10^{-3}$. The result demonstrated is obtained from fixing the random number generator seed. Figure 6 shows the testing results for a randomly chosen sample. All models are trained on an RTX 3090 or an A4000. The codes to replicate the experiments are open-source and publicly available. [1].

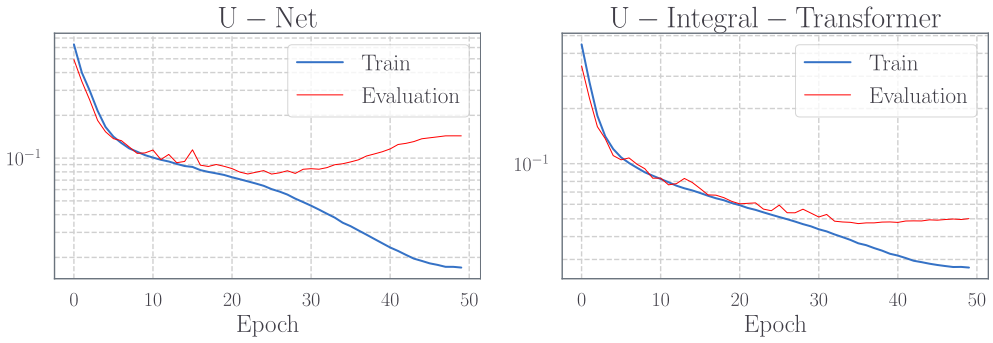

Figure 4: The left: the training-testing pixel-wise binary cross entropy convergence for the CNN-based U-Net with 31 million parameters, a clear overfitting pattern is shown. The right: the training-testing convergence for the attention-based U-Transformer with 11.4 million parameters.

C.2 Network Architecture

The difference in architectural hyperparameters, together with training and evaluation costs comparison for all the models compared in the task of EIT reconstruction can be found in Table 3.

U-Integral Transformer

- *Overall architecture.* The U-Integral-Transformer architecture is a drop-in replacement of the standard CNN-based U-Net baseline model (7.7m) in Table 1. One of the novelties is that the input is a tensor that concatenates different measurement matrices as different channels, and a similar practice can be found in Brandstetter et al. (2023). Same with the baseline U-Net, the UIT has three downsampling layers as the encoder (feature extractor). The downsampling layers map $m \times m$ latent representations to $m/2 \times m/2$, and expand the number of channels from C to $2C$. To leverage the "basis $\Leftrightarrow$ channel" interpretation and the basis update nature of the attention

[1] https://github.com/scaomath/eit-transformer

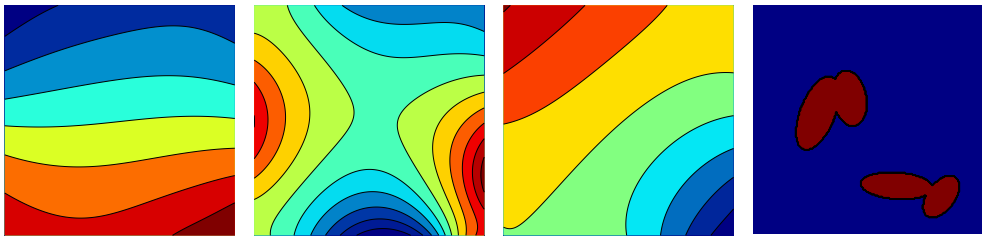

Figure 5: The harmonic extension feature map ϕ_l (left 1-3 as different channels' inputs to the neural network) corresponding to a randomly chosen sample's inclusion map (right). No visible relevance shown with the ground truth. The layered heatmap appearance is adopted in plotly contour only for aesthetics purposes, no edge-like nor layer-like features can be observed from the actual harmonic extension feature map input.

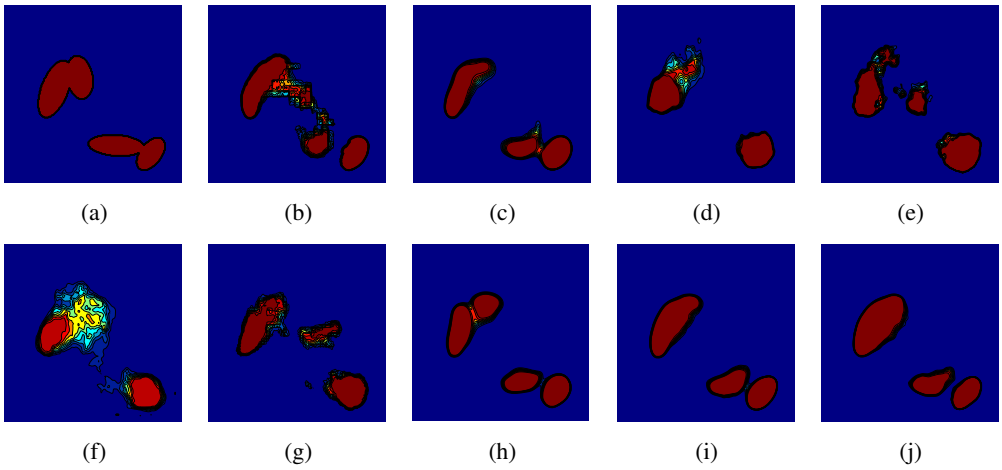

Figure 6: The neural network evaluation result for the inclusion in Figure 5 using various models. A model's input is the left most function in Figure 5 if this model uses a single channel, or all left three in Figure 5 if a model uses 3-channel input. (a) Ground truth inclusion in Figure 5; (b) U-Net baseline (7.7m) prediction; (c) U-Net big (31m) prediction with 3 channels; (d) Fourier Neural Operator (10.4m) prediction with 1 channel; (e) Fourier Neural Operator big (33m) prediction with 1 channel; (f) Adaptive Fourier Neural Operator (10.7m) prediction with 1 channel; (g) Multiwavelet Neural Operator (9.8m) prediction with 1 channel; (h) Hybrid UT with a linear attention (10.13m) prediction with 1 channel; (i) UIT (11.4m) prediction with 1 channel; (j) UIT (11.4m) prediction with 3 channels.

mechanism in Section 3.3, the proposed attention block is first added on the coarsest grid, which has the most number of channels. UIT has three upsampling layers as the decoder (feature selector), which map $m/2 \times m/2$ latent representations to $m \times m$, and shrink the number of channels from $2C$ to C. In these upsampling layers, attention blocks are applied on each cross-layer propagation to compute the interaction between the latent representations on both coarse and fine grids (see below). Please refer to Figure 8 for a high-level encoder-decoder schematic.

- *Double convolution block.* The double convolution block is modified from that commonly seen in Computer Vision (CV) models, such as ResNet (He et al., 2016). We modify this block such that upon being used in an attention block, the batch normalization (Ioffe & Szegedy, 2015) can be replaced by the layer normalization (Ba et al., 2016), which can be understood as a learnable approximation to the Gram matrices' inverse by a diagonal matrix.
- *Positional embedding.* At each resolution, the 2D Euclidean coordinates of an $m \times m$ regular Cartesian grid are the input of a channel expansion through a fixed learnable linear layer and are then added to each latent representation. This choice of positional embedding enables a bilinear interpolation between the coarse and fine grids or vice versa (see below).

- *Mesh-normalized attention.* The scaled dot-product attention in the network is chosen to be the integral kernel attention in (11) with a mesh-based normalization. Please refer to Figure 7 for a diagram in a single attention head.

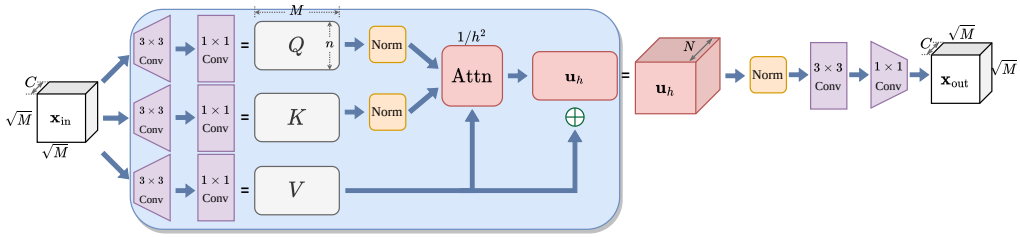

Figure 7: Detailed flow of the modified 2D attention-based encoder layer using (11). C: the number of channels in the input, N: the number of expanded channels (for the basis expansion interpretation in Theorem 2).

- *Interpolation.* Instead of max pooling used in the standard CNN-based U-Net, we opt for a bilinear interpolation on the Cartesian grid to map a latent representation from the fine grid to the coarse grid or vice versa. Note that, in the upsampling layers, the interpolation's outputs are directly inputted into an attention block that computes the interaction of latent representations between coarse and fine grids (see below).
- *Coarse-fine attention in up blocks.* A modified attention in Section 3.3 with a pre-inner-product normalization replaces the convolution layer on the coarsest level. The skip connection from the encoder latent representations to the ones in the decoder are generated using an architecture similar to the cross attention used in Petit et al. (2021). Q and K are generated from the latent representation functions on the same coarser grids. As such, the attention kernel to measure the interaction between different channels is built from the coarse grid. V is associated with a finer grid. Compared with the one in Petit et al. (2021), the modified attention in our method is inspired by the kernel integral for a PDE problem. Thus, it has (1) no softmax normalization or (2) no Hadamard product-type skip connection.

Other operator learners compared

- *Fourier Neural Operator (FNO) and variants.* Fourier Neural Operator (FNO2d) learns convolutional filters in the frequency domain for some pre-selected modes, efficiently capturing globally-supported spatial interactions for these modes. The weight filter in the frequency domain multiplies with the lowest modes in the latent representations (four corners in the FFT). The Adaptive Fourier Neural Operator (AFNO2d) adds a token-mixing layer, as seen in Figure 2 in Guibas et al. (2022), appending every spectral convolution layer in the baseline FNO2d model.
- *Multi-Wavelet Neural Operator (MWO).* The MultiWavelet Neural Operator (MWO) is proposed in Gupta et al. (2021), which introduces a multilevel structure into the FNO architecture. MWO still follows FNO's practice on each level by pre-selecting the lowest modes.

D From DSM to Transformer

This section gives a more detailed presentation of how the attention-like operator is derived from (15) with learnable kernels. We begin with recalling the original indicator function from (10):

$$\mathcal{I}^D(x) := C_{f,g} \frac{\mathbf{d}(x) \cdot \nabla \phi(x)}{|\eta_x|_{H^s(\partial\Omega)}}, \tag{30}$$

where $C_{f,g} = \|f - \Lambda_{\sigma_0} g\|^{-1}_{L^2(\partial\Omega)}$ is a constant, and the equations of the functions ϕ and η_x are :

$$-\Delta\phi = 0 \quad \text{in} \quad \Omega, \quad \mathbf{n} \cdot \nabla\phi = (f - \Lambda_{\sigma_0} g) + \xi \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega} \phi \, \mathrm{d}s = 0, \tag{31}$$

$$-\Delta\eta_x = -\mathbf{d}(x) \cdot \nabla\delta_x \quad \text{in} \quad \Omega, \quad \mathbf{n} \cdot \nabla\eta_x = 0 \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega} \eta_x \, \mathrm{d}s = 0. \tag{32}$$

Table 3: The detailed comparison of the networks used in this study. For U-Net-based neural networks, the channel/width is the number of the base channels on the finest grid after the initial channel expansion. A `torch.cfloat` type parameter entry counts as two parameters. GFLOPs: Giga FLOPs for 1 backpropagation (BP) performed for a batch of 8 samples recorded the PyTorch `autograd` profiler for 1 BP averaging from 100 BPs. Eval: number of instances per second.

	Architectures				Training/Evaluation cost		# params
	layers	channel/width	modes	norm	GFLOPs	eval	
U-Net	7	64	N/A	batch	140.6	298.6	7.70m
U-Net big	9	64	N/A	batch	184.1	273.4	31.04m
FNO2d	6	48	14	FFT	196.5	235.4	10.86m
AFNO2d	6	48	14	FFT	198.4	151.0	10.88m
MWO2d	4	64	12	Legendre	1059	59.6	9.81m
Hybrid UT	7	32	N/A	batch	427.5	114.0	10.13m
Cross-Attn UT	7	64	N/A	layer	658.9	103.3	11.42m
UIT (ours)	7	64	N/A	layer	658.3	104.8	11.43m

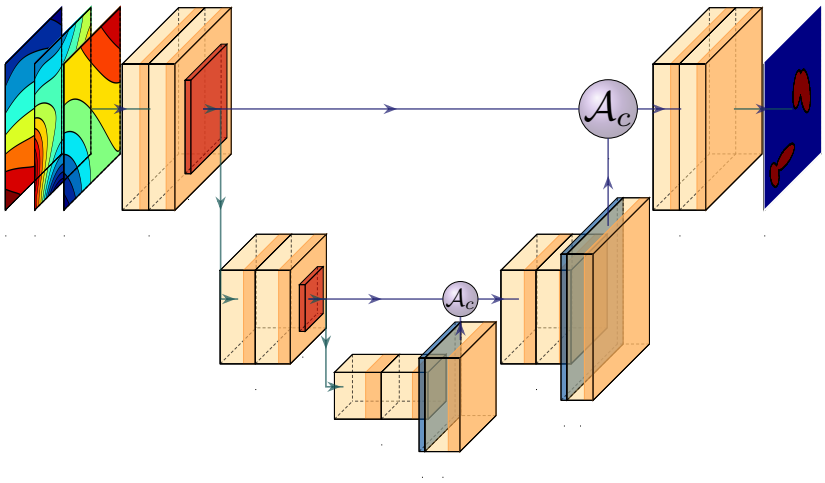

Figure 8: A simplified schematic of the U-Integral-Transformer that follows the standard U-Net. The input is a tensor concatenating the discretizations of ϕ and $\nabla\phi$. The output is the approximation to the index function $\mathcal{I}^D$. ■: 3×3 convolution + ReLU; ■: layer normalization or batch normalization; ■: bilinear interpolations from the fine grid to the coarse grid; ■: cross attention $\mathcal{A}_c$ that uses the latent representations on a coarse grid to compute interactions to produce the latent representations on a finer grid; ■: input and output discretized functions in certain Hilbert spaces. The TikZ source code to produce this figure is modified from the examples in Iqbal (2018).

Chow et al. (2014) shows that $\mathcal{I}^D(x)$ can be written as a sum of Gaussian-like distributions that attain maximum values for $x \in D$ or close to D. However, the accuracy is much limited by empirical choices of quantities in (30), for example, $\mathbf{d}(x) = \nabla\phi(x)/\|\nabla\phi(x)\|$ and $s = 3/2$. In addition, the frequency of the boundary data plays an important role in reconstruction; see the derivation in Chow et al. (2014, Section 5) for circular Ω. What is more, such a simple formula can be derived only for a single data pair.

Henceforth, these "empirical chosen quantities" are made to be learnable from data by introducing two undetermined kernels $\mathcal{K}(x, y)$, $\mathcal{Q}(x, y)$, and a self-adjoint positive definite linear operator $\mathcal{V}$, the modified indicator function is written as

$$\hat{\mathcal{I}}_1^D(x) := C_{f,g} \frac{\int_\Omega \mathbf{d}(x) \cdot \mathcal{K}(x, y) \nabla\phi(y) \, \mathrm{d}y}{|\eta_x|_Y}. \tag{33}$$

with

$$\mathbf{d}(x) := \int_\Omega \mathcal{Q}(x, y) \nabla\phi(y) \, \mathrm{d}y. \tag{34}$$

$$|\eta_x|_Y^2 := (\mathcal{V}\eta_x, \eta_x)_{L^2(\partial\Omega)}. \tag{35}$$

Applying certain quadrature rule to (33) with the quadrature points z_i, i.e., the grid points of Ω_h, and weights $\{\omega_j\}$, we obtain an approximation to the integral:

$$\int_\Omega \mathcal{K}(z_i, y)\partial_{x_n}\phi(y)\,\mathrm{d}y \approx \sum\nolimits_j \omega_j \mathcal{K}(z_i, z_j)\partial_{x_n}\phi(z_j) =: \mathbf{k}_i^T \boldsymbol{\phi}_n, \tag{36}$$

i.e., $\mathbf{q}_i^T$ is the vector of $[\omega_j \mathcal{K}(z_i, z_j)]_j$. For (34), we consider one component $d_n(x)$ of $\mathbf{d}(x)$. With the same rule to compute the integral, (16) can be written as

$$d_n(z_i) \approx \sum\nolimits_j \omega_j \mathcal{Q}(z_i, z_j)\partial_{x_n}\phi(z_j) =: \mathbf{q}_i^T \boldsymbol{\phi}_n, \tag{37}$$

i.e., $\mathbf{q}_i^T$ is the vector of $[\omega_j \mathcal{Q}(z_i, z_j)]_j$. Next, we proceed to express $|\eta_{z_i}|_Y$ by discretizing the variational form in (32) using a linear finite element method (FEM). Applying integration by parts, the weak form of (32) at $x = z_i$ is

$$\int_\Omega \nabla\eta_{z_i} \cdot \nabla\psi\,\mathrm{d}y = \int_\Omega -\mathbf{d}(z_i)\cdot\nabla\delta_{z_i}(y)\psi(y)\,\mathrm{d}y = -\mathbf{d}(z_i)\cdot\nabla\psi(z_i), \tag{38}$$

for any test function $\psi \in H_0^1(\Omega)$. Here, we let $\{\psi^{(j)}\}_{j=1}^M$ be the collection of the finite element basis functions, and for the fixed z_i we further let $\boldsymbol{\psi}_{n,i}$ be the vector approximating $[\partial_{x_n}\psi^{(j)}(z_i)]_{j=1}^M$. Denote $\boldsymbol{\eta}_i$ as the vector approximating $\{\eta_{z_i}(z_j)\}_{j=1}^M$. Introduce the matrices

- B: the finite element/finite difference discretization of $-\Delta$ on Ω_h (a discrete Laplacian) coupled with the Neumann boundary condition and the zero integral normalization condition in (9).
- R: the matrix that projects a vector defined at interior grids to the one defined on $\partial\Omega$.

Then, the finite element discretization of (38) yields the linear system:

$$B\boldsymbol{\eta}_i = \sum\nolimits_n d_n(z_i)\boldsymbol{\psi}_{n,i} \approx \sum\nolimits_n \boldsymbol{\psi}_{n,i}\mathbf{q}_i^T\boldsymbol{\phi}_n, \tag{39}$$

where we have used (37). Then, the trace of η_{z_i} on $\partial\Omega$ admits the following approximation

$$\bar{\boldsymbol{\eta}}_i := \eta_{z_i}|_{\partial\Omega} \approx \sum\nolimits_n RB^{-1}\boldsymbol{\psi}_{n,i}\mathbf{q}_i^T\boldsymbol{\phi}_n. \tag{40}$$

Now, we can discretize (35). Note that the trace of a linear finite element space on $\partial\Omega$ is still a continuous piecewise linear space, defined as $S_h(\partial\Omega)$. Then, the self-adjoint positive definite operator $\mathcal{V}$ can be parameterized by a symmetric positive definite (SPD) matrix denoted by V operating on the space $S_h(\partial\Omega)$. We can approximate $|\eta_{z_i}|_Y^2$ as

$$|\eta_{z_i}|_Y^2 \approx \bar{\boldsymbol{\eta}}_i^T V \bar{\boldsymbol{\eta}}_i \approx \sum\nolimits_n \boldsymbol{\psi}_{n,i}^T B^{-1}R^T V R B^{-1}\boldsymbol{\psi}_{n,i}\boldsymbol{\phi}_n^T\mathbf{q}_i\mathbf{q}_i^T\boldsymbol{\phi}_n \tag{41}$$

where $\boldsymbol{\psi}_{n,i}^T B^{-1}R^T V R B^{-1}\boldsymbol{\psi}_{n,i} \geq 0$ as V is SPD. Define

$$\mathbf{v}_{n,i} = (\boldsymbol{\psi}_{n,i}^T B^{-1}R^T V R B^{-1}\boldsymbol{\psi}_{n,i})^{1/2}\mathbf{q}_i. \tag{42}$$

This can be considered another learnable vector since the coefficient of $\mathbf{q}_i$ comes from the learnable matrix V. Then, (41) reduces to

$$|\eta_{z_i}|_Y^2 \approx \sum_n \boldsymbol{\phi}_n^T\mathbf{v}_{n,i}\mathbf{v}_{n,i}^T\boldsymbol{\phi}_n \tag{43}$$

Putting (36), (37) and (43) into (15), we have

$$\hat{\mathcal{I}}_1^D(z_i) \approx \Big\{\|f - \Lambda_{\sigma_0}g\|_{L^2(\partial\Omega)}^{-1}\big(\sum_n \boldsymbol{\phi}_n^T\mathbf{v}_{n,i}\mathbf{v}_{n,i}^T\boldsymbol{\phi}_n\big)^{-1/2}\Big\}\sum_n \boldsymbol{\phi}_n^T\mathbf{q}_i\mathbf{k}_i^T\boldsymbol{\phi}_n. \tag{44}$$

Now, using the notation in (21), we get the desired representation (22).

E PROOF OF THEOREM 1

Lemma 1. *Suppose the boundary data g_l is the eigenfunction of $\Lambda_\sigma - \Lambda_{\sigma_0}$ corresponding to the l-th eigenvalue λ_l, and let ϕ_l be the data functions generated by harmonic extensions*

$$-\Delta\phi_l = 0 \quad in \quad \Omega, \quad \mathbf{n}\cdot\nabla\phi_l = (f_l - \Lambda_{\sigma_0} g_l) = (\Lambda_\sigma - \Lambda_{\sigma_0})g_l \quad on \quad \partial\Omega, \quad \int_{\partial\Omega} \phi_l \, \mathrm{d}s = 0, \tag{45}$$

where $l = 1, 2, \cdots$. Let $\mathbf{d}$ be an arbitrary unit vector in $\mathbb{R}^2$, define a function

$$\Theta_L(x) = \sum_{l=1}^{L} \frac{(\mathbf{d}\cdot\nabla\phi_l(x))^2}{\lambda_l^3}. \tag{46}$$

Then, there holds

$$\lim_{L\to\infty} \Theta_L(x) = \begin{cases} \infty, & \text{if } x \notin D, \\ \text{a finite constant}, & \text{if } x \in D. \end{cases} \tag{47}$$

Proof. See Theorem 4.1 in Guo & Jiang (2020), and also see Brühl (2001); Hanke & Brühl (2003). □

Theorem 1 (A finite-dimensional approximation of the index function)**.** *Suppose the boundary data g_l is the eigenfunction of $\Lambda_\sigma - \Lambda_{\sigma_0}$ corresponding to the l-th eigenvalue λ_l, and let ϕ_l be the data functions generated by harmonic extensions given in (9). Define the space:*

$$\widetilde{\mathbb{S}}_L = \text{Span}\{\partial_{x_1}\phi_l \; \partial_{x_2}\phi_l : \; l = 1, ..., L\}, \tag{48}$$

and the dictionary:

$$\mathbb{S}_L = \{a_1 + a_2 \arctan(a_3 v) \; : \; v \in \widetilde{\mathbb{S}}_L, \; a_1, a_2, a_3 \in \mathbb{R}\}. \tag{49}$$

Then, for any $\epsilon > 0$, we con construct an index function $\mathcal{I}_L^D \in \mathbb{S}_L$ s.t.

$$\sup_{x\in\Omega} |\mathcal{I}^D(x) - \mathcal{I}_L^D(x)| \le \epsilon \tag{50}$$

provided L is large enough.

Proof. Consider the function $\Theta_L(x)$ from Lemma 1. As $\Theta_L(x) > 0$, it is increasing with respect to L. Then, there is a constant ρ such that $\rho > \Theta_L(x)$, $\forall x \in D$. Given any $\epsilon > 0$, there is an integer L such that $\Theta_L(x) > 4\rho\epsilon^{-2}/\pi^2$, $\forall x \notin D$. Define

$$\mathcal{I}_L^D(x) = 1 - \frac{2}{\pi}\arctan\left(\frac{\pi\epsilon}{2\rho}\Theta_L(x)\right) \tag{51}$$

Note the fundamental inequality $z > \arctan(z) \ge \frac{\pi}{2} - z^{-1}$, $\forall z > 0$. Then, if $x \in D$, there holds

$$|\mathcal{I}^D(x) - \mathcal{I}_L^D(x)| = \frac{2}{\pi}\arctan\left(\frac{\pi\epsilon}{2\rho}\Theta_L(x)\right) < \frac{\epsilon}{\rho}\Theta_L(x) < \epsilon$$

if $x \notin D$, there holds

$$|\mathcal{I}^D(x) - \mathcal{I}_L^D(x)| = 1 - \frac{2}{\pi}\arctan\left(\frac{\pi\epsilon}{2\rho}\Theta_L(x)\right) \le \frac{4\rho}{\pi^2\epsilon\Theta_L(x)} < \epsilon.$$

Therefore, the function in (51) fulfills (50).

□

F PROOF OF THEOREM 2

In presenting Theorem 2 in Section 3.3, we use the term "multiplicative" to describe the fact that two latent representations are multiplied in the attention mechanism. In contrast, no such operation exists in, e.g., a pointwise FFN or a convolution layer. Heuristically speaking, the main result in Theorem 2 states that the output latent representations can be of a higher "frequency" than the input if the neural network architecture has "multiplicative" layers in it. The input latent representations are discretizations of certain functions, and they are combined using matrix dot product as the one used in attention. Suppose that this discretization can represent functions of such a frequency with a certain approximation error, the resulting matrix/tensor can be an approximation of a function with a higher frequency than the existing latent representation under the same discretization. Please see Figure 9 and Figure 10 for empirical evidence of this phenomenon for the latent representations: with completely smooth input (harmonic extensions), e.g., see Figure 5, attention-based learner can generate latent representations with multiple peaks and valleys.

Theorem 2 (Frequency-bootstrapping for multiplicative neural architectures)**.** *Consider $\Omega = (0, \pi)$ which has a uniform discretization of $\{z_i\}_{i=1}^M$ of size h, and $v(x) = \sin(ax)$ for some $a \in \mathbb{Z}^+$. Let $N := a - 1 \geq 1$ be the number of channels in the attention layer of interest, assume that (i) the current latent representation $\mathbf{p}_h \in \mathbb{R}^{M\times N}$ consists of the discretization of the first N Legendre polynomials $\{p_j(\cdot)\}_{j=1}^N$ such that $(\mathbf{p}_h)_{ij} = p_j(z_i)$, (ii) $\mathbf{p}_h$ is normalized and the normalization weights $\alpha \equiv 1$ in (14), (iii) the discretization satisfies that $|\sum_{i=1}^M hf(z_i) - \int_\Omega f(x)\,\mathrm{d}x| \leq C\,h$. Then, there exists a set of attention weights $\{W^Q, W^K, W^V\}$ such that for $u(x) = \sin(a'x)$ with $\mathbb{Z}^+ \ni a' > a$*

$$\|\tilde{u} - u\|_{L^2(\Omega)} \leq C \max\{h, \|\varepsilon\|_{L^\infty(\Omega)}\}, \tag{52}$$

where $\tilde{u}$ and $\tilde{\kappa}(\cdot,\cdot)$ are defined as the output of and the kernel of the attention formulation in (14), respectively; $\varepsilon(x) := \|\kappa(x,\cdot) - \tilde{\kappa}(x,\cdot)\|_{L^2(\Omega)}$ is the error function for the kernel approximation.

Proof. Without loss of generality, it is assumed that $a' = a + 1$. The essential technical tools used suggest the validity for any $a' > a > 0$ ($a', a \in \mathbb{Z}^+$). Consider a simple non-separable smooth kernel function

$$\kappa(x, z) := \sin((a + 1)(x - z)), \tag{53}$$

it is straightforward to verify that for $v(x) := \sin(ax)$, we have $c_1 = 2a/(2a + 1)$

$$\int_\Omega \kappa(x, z)v(x)\,\mathrm{d}x = \int_\Omega \sin\big((a + 1)(x - z)\big)\sin(ax)\,\mathrm{d}x = c_1 \sin\big((a + 1)z\big) =: c_1 u(z). \tag{54}$$

As is shown, it suffices to show that the matrix multiplication (a separable kernel) in the attention mechanism approximates this non-separable kernel with an error related to the number of channels. To this end, taking the Taylor expansion, centered at a $z_0 \in \Omega$ of $\kappa(x,\cdot)$ with respect to the second variable at each $x \in \Omega$, we have

$$\kappa_N(x, z) := \sum_{l=1}^N \frac{(z - z_0)^{l-1}}{(l - 1)!} \frac{\partial^{l-1}\kappa}{\partial z^{l-1}}(x, z_0). \tag{55}$$

It is straightforward to check that

$$\|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} \leq c_1 \frac{\sqrt{(\pi - z_0)^{2N+1} + z_0^{2N+1}}}{N!\sqrt{2N + 1}} \left\|\frac{\partial^N \kappa}{\partial z^N}(x,\cdot)\right\|_{L^\infty(\Omega)}. \tag{56}$$

By the assumptions on $\kappa(\cdot,\cdot)$, and a straightforward computation we have

$$\|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} \leq c_2 \frac{(\pi)^N}{N!\sqrt{N}} \left\|\frac{\partial^N \kappa}{\partial z^N}(x,\cdot)\right\|_{L^\infty(\Omega)}. \tag{57}$$

Next, let $q_l(z) := (z - z_0)^{l-1}/(l - 1)!$ and $k_l(x) := \partial^{l-1}\kappa/\partial z^{l-1}(x, z_0)$ for $1 \leq l \leq N$, i.e., they form a Pincherle-Goursat (degenerate) kernels (Kress, 1999, Chapter 11)

$$\kappa_N(x, z) = \sum_{l=1}^N q_l(x)k_l(z). \tag{58}$$

By this choice of the latent representation space being the first N Legendre polynomials, $q_l \in \mathbb{Y} := \mathrm{span}\{p_j\}$, thus there exists a set of weights $\{w_l^Q \in \mathbb{R}^l\}_{l=1}^N$ corresponding to each channel, such that

$$q_l(\cdot) = \sum_{j=1}^N w_{l,j}^Q p_j(\cdot) =: \tilde{q}_l(\cdot). \tag{59}$$

This is to say, $Q = \mathbf{p}_h W^Q \in \mathbb{R}^{M\times N}$ with the l-th column of Q being the discretization of $\tilde{q}_l(\cdot)$.

For the key matrix, by standard polynomial approximation since $\mathbb{Y} \simeq \mathbb{P}_N(\Omega)$, there exists a set of weights $\{w_l^K \in \mathbb{R}^N\}_{l=1}^N$, such that

$$k_l(\cdot) \approx \sum_{j=1}^N w_{l,j}^K p_j(\cdot) =: \tilde{k}_l(\cdot), \tag{60}$$

i.e., $K = \mathbf{p}_h W^K \in \mathbb{R}^{M\times N}$ with the l-th column of K being the discretization of $\tilde{k}_l(\cdot)$. Moreover, it can approximate $k_l(\cdot)$ with the following estimate

$$\|k_l(\cdot) - \tilde{k}_l(\cdot)\|_{L^2(\Omega)} \le c_3 \frac{\pi^N}{2^N(N+1)^N} |k_l(\cdot)|_{H^N(\Omega)}. \tag{61}$$

Similarly, without loss of generality, we choose $v(\cdot) := v_1(\cdot)$, which is concatenated to V such that it occupies the first channel of V defined earlier, we have $\{w_l^V \in \mathbb{R}^N\}$ such that

$$v(\cdot) \approx \sum_{j=1}^N w_{1,j}^K p_j(\cdot) =: \tilde{v}(\cdot) \tag{62}$$

and

$$\|v(\cdot) - \tilde{v}(\cdot)\|_{L^2(\Omega)} \le c_4 \frac{\pi^N}{2^N(N+1)^N} |v(\cdot)|_{H^N(\Omega)}. \tag{63}$$

Now, to approximate the frequency bootstrapping in (54), define

$$\tilde{u}(z) := \int_\Omega \tilde{\kappa}_N(x,z)\tilde{v}(x)\,\mathrm{d}x, \quad \text{with } \tilde{\kappa}_N(x,z) := \sum_{l=1}^N \tilde{q}_l(x)\tilde{k}_l(z). \tag{64}$$

Then, we have for any $z \in \Omega$

$$\begin{aligned} u(z) - \tilde{u}(z) &= \int_\Omega \kappa(x,z)v(x)\,\mathrm{d}x - \int_\Omega \kappa_N(x,z)\tilde{v}(x)\,\mathrm{d}x \\ &= \int_\Omega \big(\kappa(x,z) - \tilde{\kappa}_N(x,z)\big)v(x)\,\mathrm{d}x + \int_\Omega \kappa_N(x,z)\big(v(x) - \tilde{v}(x)\big)\,\mathrm{d}x. \end{aligned} \tag{65}$$

Thus, we have

$$\begin{aligned} \|u - \tilde{u}\|_{L^\infty(\Omega)} &\le \max_{z\in\Omega}\left\{\int_\Omega \left|\big(\kappa(x,z) - \tilde{\kappa}_N(x,z)\big)v(x)\right|\,\mathrm{d}x + \int_\Omega \left|\kappa_N(x,z)\big(v(x) - \tilde{v}(x)\big)\right|\,\mathrm{d}x\right\} \\ &\le \max_{z\in\Omega}\Big\{\underbrace{\|\kappa(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)}}_{(*)}\|v\|_{L^2(\Omega)} + \|\tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)}\|v - \tilde{v}\|_{L^2(\Omega)}\Big\}. \end{aligned} \tag{66}$$

Now, by triangle inequality, $q_l = \tilde{q}_l$, the definitions above (53) implying $|k_l(\cdot)|_{H^N(\Omega)} \le c\,2^N$ and $\|q_l\|_{L^2(\Omega)} \le c\pi^N/(\sqrt{N}N!)$, and the estimate in (61)

$$\begin{aligned} (*) &\le \|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} + \|\kappa_N(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)} \\ &\le \|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} + \sum_{l=1}^N \|q_l\|_{L^2(\Omega)}\|k_l - \tilde{k}_l\|_{L^2(\Omega)} \\ &\le c_5\left(\frac{(2\pi)^N}{N!\sqrt{N}} + c_6\frac{\pi^N}{2^N(N+1)^N}\sum_{l=1}^N \|q_l\|_{L^2(\Omega)}|k_l(\cdot)|_{H^N(\Omega)}\right) \\ &\le c_6\frac{(2\pi)^N}{N!\sqrt{N}}. \end{aligned} \tag{67}$$

Notice this is the same order with the estimate of $\|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)}$. For the term $\|\tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)}$, a simple triangle inequality trick can be used:

$$\begin{aligned}\|\tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)} &\leq \|\kappa_N(x,\cdot)\|_{L^2(\Omega)} + \|\kappa_N(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)} \\ &\leq \|\kappa(x,\cdot)\|_{L^2(\Omega)} + \|\kappa_N(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)},\end{aligned} \tag{68}$$

which can be further estimated by reusing the argument in (67).

Lastly, using the following argument and the estimate for $\|u - \tilde{u}\|_{L^\infty(\Omega)}$ yield the desired result:

$$\|u - \tilde{u}\|^2_{L^2(\Omega)} \leq \|u - \tilde{u}\|_{L^1(\Omega)} \|u - \tilde{u}\|_{L^\infty(\Omega)} \leq 2 \max |u| \, \|u - \tilde{u}\|^2_{L^\infty(\Omega)}. \tag{69}$$

□

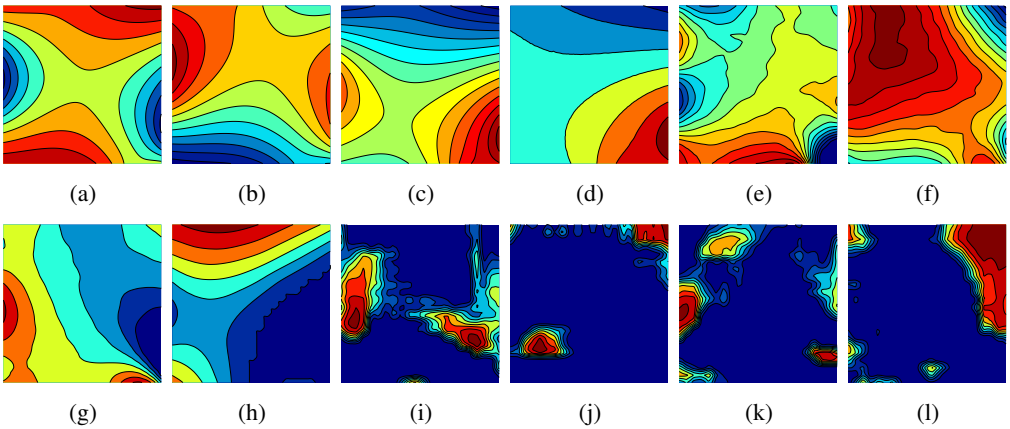

Figure 9: The latent representations from CNN-based UNet when evaluating for the sample in Figure 5, in their respective layers, 4 latent representations are extracted from 4 randomly selected channels: (a)–(d) are from the feature extracting layer (layer 1, 128×128 grid), (e)–(h) are from the middle layer acting on the coarsest level (32×32 grid), (i)–(l) are from the next level (64×64 grid).

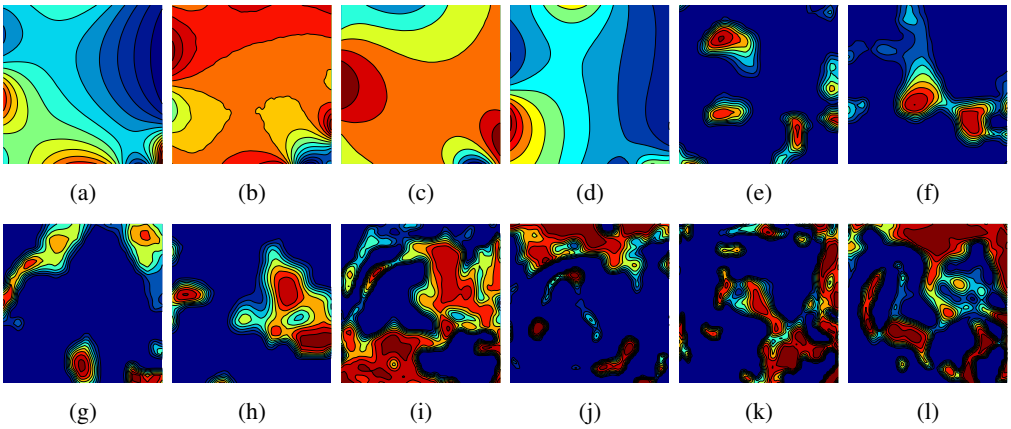

Figure 10: The latent representations from UIT when evaluating for the sample in Figure 5, in their corresponding layers with respect to those in Figure 9, 4 latent representations are extracted from 4 randomly selected channels: (a)–(d) are from the feature extracting layer (layer 1, 128×128 grid), (e)–(h) are from the middle layer acting on the coarsest level (32×32 grid), (i)–(l) are from the next level (64×64 grid).

G LIMITATIONS, EXTENSIONS, AND FUTURE WORK

In this study, the σ to be recovered relies on a piecewise constant assumption. This assumption is commonly seen in the theoretical study of the original DSM. For many EIT applications in medical imaging and industrial monitoring, σ may involve non-sharp transitions or even contain highly anisotropic/multiscale behaviors making it merely an L^∞ function. If the boundary data pairs are still quite limited, i.e., only a few electric modes are placed on the boundary $\partial\Omega$, the proposed model alone is not expected to perform as well as in benchmark problems. Nevertheless, it can still contribute to achieving reconstruction with satisfactory accuracy, if certain a priori knowledge of the problem is accessible. End2end-wise, our proposed method has limitations like other operator learners: the data manifold on which the operator is learned is assumed to exhibit low-dimensional/low-rank attributes. The behavior of the operator of interest on a compact subset is assumed to be reasonably well approximated by a finite number of bases. Therefore, for non-piecewise constant conductivities, the modification can be to employ a suitable data set, in which the sampling of $\{\sigma^{(k)}\}$ represents the true σ's distribution a posteriori to a certain degree. However, to reconstruct non-piecewise constant conductivities, more boundary data pairs or even the entire NtD map is demanded from a theoretical perspective (Astala & Päivärinta, 2006a;b; Nachman, 1996; Kohn & Vogelius, 1984; Sylvester & Uhlmann, 1987). For fewer data pairs and more complicated conductivity set-up, there have been efforts in this direction hierarchically using matrix completion (Bui-Thanh et al., 2022) to recover Λ_σ. When Λ_σ is indeed available, σ can be described by a Fredholm integral equation, see Nachman (1996, Theorem 4.1), which itself is strongly related to the modified attention mechanism (14) of the proposed Transformer. The architectural resemblance may lead to future explorations in this direction. Optimization with regularization can be applied for the instance of interest (fine-tune) from the perspective of improving the reconstruction for a single instance. This approach dates back to the classical iterative methods involving adaptivity (Jin & Xu, 2019). Recent novel DL-inspired adaptions (Li et al., 2021b; Benitez et al., 2023) re-introduce this type of method. In fine-tuning, the initial guess is the reconstruction by the operator learner trained in the end2end pipeline (pre-train).