TRANSFORMER MEETS BOUNDARY VALUE INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A Transformer-based deep direct sampling method is proposed for solving a class of 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 but critical question: whether and how one can benefit from the theoretical structure of a mathematical problem to develop task-oriented and structure-conforming deep neural network? Specifically, inspired by direct sampling methods for inverse problems, the 1D boundary data are preprocessed by a partial differential equation-based feature map to yield 2D harmonic extensions in different frequencies as different input channels. Then, by introducing learnable non-local kernel, the approximation of direct sampling is recast to a modified attention mechanism. The proposed method is then applied to electrical impedance tomography, a well-known severely ill-posed nonlinear inverse problem. The new method achieves superior accuracy over its predecessors and contemporary operator learners, as well as shows robustness with respect to noise. This research shall strengthen the insights that the attention mechanism, despite being invented for natural language processing tasks, 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.

Let $\mathcal{F} : \mathbb{P} \to \mathbb{M}$ be a nonlinear operator associated a physical model governed by certain partial differential equations (PDE) on a bounded Ω in $\mathbb{R}^2$. For the measurement $\mathbf{m} \in \mathbb{M}$ on the boundary $\partial\Omega$, and the coefficient $\mathbf{p} \in \mathbb{P}$ of the underlying PDEs to be recovered, the forward PDE model is

$$\mathbf{m} = \mathcal{F}(\mathbf{p}) + \boldsymbol{\xi} \tag{1}$$

where $\boldsymbol{\xi}$ is a noise following certain distribution. Comparing with the classic linear inverse problems $\mathbf{y} = A\mathbf{x} + \boldsymbol{\epsilon}$ in computer vision and signal processing, e.g. (Marroquin et al., 1987), to recover a signal $\mathbf{x}$ from measurement $\mathbf{y}$ with noise $\boldsymbol{\epsilon}$, the fundamental difference is that $\mathcal{F}$ is highly nonlinear, and the Hilbert spaces $\mathbb{M}$ and $\mathbb{P}$ are usually infinite dimensional. In many situations, even though the inverse operator $\mathcal{F}^{-1} : \mathbb{M} \to \mathbb{P}$ is theoretically well-defined, as the infinite dimensional $\mathbb{M}$ is practically approximated by a finite dimensional (sub)space $\mathbb{M}_L$, the problem of seeking the approximated operator $\mathcal{F}_L^{-1} : \mathbb{M}_L \to \mathbb{P}$ is usually highly ill-posed (not having a well-defined unique output) and poses great challenges to the reconstruction algorithms.

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., 2018; Rondi & Santosa, 2001; Chen et al., 2020). One usually looks for the desired $\mathbf{p}$ by solving a minimization problem

$$\inf_{\mathbf{p}\in\mathbb{P}} \left\{ \sum\nolimits_{m\in\mathbb{M}_L} \|\mathcal{F}(\mathbf{p}) - \mathbf{m}\|_{\mathbb{M}} \right\} + \mathcal{R}(\mathbf{p}), \tag{2}$$

where $\mathcal{R}(\mathbf{p})$ is a regularization term to alleviate the ill-posedness, and its design plays a critical role for a successful reconstruction (Tarvainen et al., 2008; Tehrani et al., 2012; Wang et al., 2012). For almost all these iterative methods, due to the ill-posedness, the computation generally takes numerous iterations to converge, and the reconstruction is highly sensitive to noise. Besides, $\mathcal{F}(\mathbf{p})$ needs to be evaluated at each iteration, which is itself expensive as it requires solving forward PDE models.

Direct methods. The second methodology is to develop a well-defined mapping $\mathcal{G}_\theta$ parametrized with θ, empirically constructed to approximate the inverse map itself:

$$\mathcal{G}_\theta \approx \mathcal{F}^{-1} : \mathbb{M} \to \mathbb{P}. \tag{3}$$

This method is referred to as non-iterative or direct methods in the literature. Distinguished from iterative approaches, direct methods are in general highly problem-specific, as they are designed on an ad hoc basis based on the mathematical structure of the various inverse operators. For instances, 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) (Chow et al., 2014; 2015; Kang et al., 2018; Ito et al., 2013; Ji et al., 2019; Harris & Kleefeld, 2019; Ahn et al., 2020) being our focus in this work. These methods generally have a closed-form expression of $\mathcal{G}_\theta$ to approximate $\mathcal{F}^{-1}$, and the parameters θ have clear mathematical meaning. 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 given $\mathbf{m}$ is distinctly fast as no training optimization is needed anymore. However, such a simple closed form of $\mathcal{G}_\theta$ admitting efficient execution may not be available in practice since some mathematical assumptions and derivation may not hold.

Boundary value inverse problems. For most cases of boundary value inverse problems in 2D, one key difference, e.g., with an image inverse problem, is that data are only available on 1D manifolds, which are used to reconstruct 2D targets. The boundary data themselves generally involve certain input-output structure, 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).

Operator learning and inverse problems. Solving an inverse problem is essentially to approximate the inverse operator $\mathcal{F}^{-1}$ but based on finitely many data. The emerging deep learning (DL) technique makes it possible to directly emulate operators. The concept resembles greatly the aforementioned classical direct methodology. But operator learners by DNNs are generally considered as black boxes. A natural question to ask is whether and how existing innovative neural architectures, usually designed for other purposes, can be modified conforming with the mathematical nature of the underlying problem, which ultimately leads to structure-conforming DNNs. Inspired by classical direct methods, we try to improve the DNN reconstruction pipeline from the architectural perspective. In this regard, the proposed study provides a positive example to a hopefully definitive answer to this question, which bridges deep learning and conventional tasks in physical sciences.

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). Some efforts have been made to couple classical reconstruction methods and CNN. Notable examples include Hamilton et al. (2019); Hamilton & Hauptmann (2018), where a CNN post-processes images obtained by the classical D-bar methods; Fan et al. (2019); Fan & Ying (2020) develop BCR-Net to mimic pseudo-differential operators appearing in many inverse problems;

a deep direct sampling method is proposed in Guo & Jiang (2020); Jiang 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 locally supported kernel whose receptive field involves only a small neighborhood of a pixel, thus layer-wise speaking, 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., 2021). FNO takes the unique 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. Others include DeepONets (Lu et al., 2021; Wang et al., 2021b; Jin et al., 2022) and Transformers (Kissas et al., 2022; Li et al., 2022; Cao, 2021).

Related studies on Transformers. The attention mechanism-based models have become the state of the art in many areas since Vaswani et al. (2017), such as language tasks and computer vision e.g., Ma et al. (2021); Dosovitskiy et al. (2021). One of the most important and attractive aspects of the attention mechanism is its unparalleled capability to model long-range interactions (Tay et al., 2021) through many efficient variants see e.g., Katharopoulos et al. (2020); Choromanski et al. (2021); Nguyen et al. (2021). The relation of the attention with kernel learning is firstly 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). Among inverse problems, Transformers have been applied in medical imaging applications including segmentation problems (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), etc. To our best knowledge, there is no work in the literature that connects 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 classical DSM, we decompose 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 Transformeris, a modified attention is able to represent target functions exhibiting higher frequency natures from lower frequency input features. In the experiments, a comparative study further demonstrates a favorable match between the Transformer and the problem structure.

3 Interplay between mathematics and neural architectures

In this section, we use EIT, a classical boundary value inverse problem, as a prominent example to walk through that the triple tensor product in the attention mechanism matches extremely well with representing a solution in the inverse operator theory of EIT. The proposed method is motivated by and in pursuit to provide examples to the answers of the following questions

(**Q1**) What is an appropriate finite dimensional data format of $\mathbf{p}$ as inputs to the neural network?

(**Q2**) Is there a suitable neural network $\mathcal{G}_\theta$ in (3) taking advantage of the mathematical structure?

3.1 From EIT to Operator Learning

The forward model of EIT is given by the following second order elliptic partial differential equation

$$\nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } \Omega, \qquad \text{where } \sigma = \sigma_1 \text{ in } D, \text{ and } \sigma = \sigma_0 \text{ in } \Omega \backslash \overline{D}. \tag{4}$$

Given the domain $\Omega \subseteq \mathbb{R}^2$, the goal is to identify the unknown inclusion D buried in Ω. The value σ_0 and σ_1 are two (approximately) known constants discontinuous across the boundary ∂D. Without loss of generality, we assume $\sigma_1 > \sigma_0$. Then, the coefficient of (4) to be recovered can be described by a characteristic function $\mathbf{p} = \mathcal{I}^D(x)$ defined for each point $x = (x_1, x_2) \in \Omega$, and $\mathcal{I}^D(x)$ is 1 for

$x \in D$ and 0 for $x \in \Omega \backslash \overline{D}$, or equivalently $\mathbf{p} = \sigma_1 \mathcal{I}^D + \sigma_0(1 - \mathcal{I}^D)$. We refer readers to Figure 2 for the discretization procedure, where $\mathbf{p}_h$ is a vector of values at the grid points approximating $\mathcal{I}^D$.

In the application of EIT, the boundary measurement space $\mathbb{M}$ contains information of how the electric potential u behaves on the boundary. By exerting a current g on the boundary, solving (4) for a unique u with this specific boundary information $\sigma \nabla u \cdot \mathbf{n}|_{\partial\Omega} = g$ (Neumann boundary condition) reveals u's behavior on the whole domain Ω. However, what can be measured is only the voltages $u|_{\partial\Omega}$ on the boundary. This procedure is called Neumann-to-Dirichlet (NtD) mapping:

$$\Lambda_\sigma : g = \sigma \nabla u \cdot \mathbf{n}|_{\partial\Omega} \in H^{-1/2}(\partial\Omega) \mapsto f = u|_{\partial\Omega} \in H^{1/2}(\partial\Omega). \tag{5}$$

The H^s-H^{-s} duality for Sobolev spaces here is for formality, and we refer the readers to Appendix A for more detailed descriptions of function spaces. One measurement is a single data pair (g, f), and all the measurements form this infinite-dimensional space:

$$\mathbb{M} := H^{-1/2}(\partial\Omega) \times \Lambda_\sigma H^{-1/2}(\partial\Omega) = \{(g, f) : g \in H^{-1/2}(\partial\Omega), f = \Lambda_\sigma g\}.$$

If the NtD mapping Λ_σ is known, the theoretical uniqueness of $\mathcal{I}^D$ can be established (Brühl, 2001; Hanke & Brühl, 2003; Astala & Päivärinta, 2006; Nachman, 1996). Mathematically, it means if one is able to measure all the current-to-voltage pairs $\{g_l, f_l = \Lambda_\sigma g_l\}_{l=1}^\infty$ in $\mathbb{M}$, where $\{g_l\}_{l=1}^\infty$ form a set of basis functions of the Hilbert space $H^{-1/2}(\partial\Omega)$, then $\mathcal{F}^{-1} : \mathbb{M} \to \mathbb{P}$, which is essentially $\Lambda_\sigma \mapsto \sigma$, is well-defined admitting explicit format (Brühl, 2001).

We shall illustrate how the mathematical setup above is realized in a discrete level. For each fixed σ, the NtD mapping in (5) can be expressed as

$$\mathbf{f} = \mathbf{A}_\sigma \mathbf{g}, \tag{6}$$

where $\mathbf{g}$ and $\mathbf{f}$ are tensors containing boundary data at grid points on the boundary. If σ is known, then (6) is obtained through a typical PDE solver. For unknown σ, using a significantly large number of pairs of $\mathbf{g}$-$\mathbf{f}$ can obtain a huge matrix $\mathbf{A}_\sigma$ approximating to Λ_σ on a reasonably fine grid. The mechanism here essentially results in a *tensor2tensor* mapping/operator from $\mathbf{A}_\sigma$ to the images of σ. In this case, the operator $\mathcal{F}^{-1}$ can be learned through a large number of σ-$\mathbf{A}_\sigma$ data pairs. In particular, the BCR-Net (Fan & Ying, 2020) is a DNN approximation falls into this category.

However, as infinity or large number of boundary data pairs is not attainable in practice, the problem of more practical interests is to use only a few data pairs $\{(g_l, f_l)\}_{l=1}^L$ for reconstruction with L being very small not enough to get any reasonably good approximation of $\mathbf{A}_\sigma$ in (6). Let $\mathbb{M}_L$ be the subspace $\mathbb{M}$ spanned by $\{(g_l, f_l)\}_{l=1}^L$. Here, the highly nonlinear forward operator for finitely many pairs $\mathcal{F} = \mathcal{F}_L : \sigma \mapsto (g_l, f_l)_{l=1}^L$ is still trivially well-defined. However, it becomes a long-standing open theoretical problem whether its inverse $\mathcal{F}_L^{-1}$ is well-defined. Namely, it is unclear if the inclusion D can be exactly identified by finite measurement on boundary.

As we assume $\sigma_1 > \sigma_0$, 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). The following theorem justifies the practice of approximating $\mathcal{I}^D$ when L is large enough. Roughly speaking, as $\mathcal{F}_\infty^{-1}$ is well defined, when $L \gg 1$, we can define an $\mathcal{F}_L^{-1}$ to approximate $\mathcal{F}_\infty^{-1}$.

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*

$$-\Delta\phi_l = 0 \quad \text{in} \quad \Omega, \quad \mathbf{n} \cdot \nabla\phi_l = (f_l - \Lambda_{\sigma_0} g_l) = (\Lambda_\sigma - \Lambda_{\sigma_0}) g_l \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega} \phi_l \, \mathrm{d}s = 0, \tag{7}$$

where $l = 1, 2, \ldots$. Define $\mathbb{S}_L = \mathrm{Span}\{\partial_{x_1}\phi_l \; \partial_{x_2}\phi_l : \; l = 1, \ldots, L\}$. Then, for any $\epsilon > 0$, there exists a sufficiently large L such that $\sup_{x\in\Omega} \inf_{\mathcal{I}_L^D \in \mathbb{S}_L} |\mathcal{I}^D(x) - \mathcal{I}_L^D(x)| \le \epsilon$.

The proof of Theorem 1 can be found in Appendix C. Nevertheless, the constructional proof of the existence of $\mathcal{I}_L^D$ still relies on the entire NtD mapping Λ_σ, which again resorts to infiniteness, thus inaccessible in real applications. Another valuable corollary of Theorem 1 is to justify using the data functions ϕ_l as the input to neural networks, see Section 3.3 for more details.

Operator learning problems for EIT. Many works have made efforts to derive a well-defined operator that estimates σ instead of an exact or accurate reconstruction (Brühl, 2001; Chow et al., 2014; Dunlop & Stuart, 2015) for even $L = 1$. Such effort could benefit modern operator learning methods. To this end, 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 discretizations for 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 coefficient σ or equivalently D.* N samples of D with different shapes and locations following certain distribution are used approximate $\mathbb{P}$. N is usually large enough to represent field applications of interest. For the k-th sample of D, we let $\mathbf{p}_h^{(k)}$ be a discretization of $\mathcal{I}^D$ on Ω_h. Accordingly, $\mathbf{m}_h^{(k)}$ is a data function associated with $\mathcal{F}_L(\mathbf{p}_h^{(k)})$.

(3) *Sampling of NtD mapping.* The most highlighted approximation is, for each sample of D, there are L input-output data pairs to sample the NtD mapping Λ_σ, which are used to generate L channels in an input image. Despite a "sufficiently large" L needed in Theorem 1, in practice, 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 $\mathcal{F}^{-1}$ by minimizing

$$\mathcal{J}(\theta) := \frac{1}{N} \sum_{k=1}^N \|\mathcal{G}_\theta(\mathbf{m}_h^{(k)}) - \mathbf{p}_h^{(k)}\|_{\mathbb{P}}^2. \tag{8}$$

The parameters N, h, L will affect the approximation in the following way: h determines the resolution to approximate D; N determines the representativity of the training data set; L determines the known spectral information of Λ_σ.

3.2 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. Solutions to many physical problems have this form, as it aggregates the interactions of a class of functions at different locations to characterize how the operator mapping behaves between infinite-dimensional spaces. More interestingly, in most applications, 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 considered problem.

In a single head, let the input of the attention block in an encoder layer be $\mathbf{s}_h \in \mathbb{R}^{M\times N}$, where $N \gg L$ is the number of channels for the latent representations, then the query Q, key K, value V are generated by three learnable projection matrices $W^Q, W^K, W^V \in \mathbb{R}^{N\times N}$: $Q = \mathbf{s}_h W^Q$, $K = \mathbf{s}_h W^K$, $V = \mathbf{s}_h W^V$. A modified dot-product attention is proposed as follows:

$$U = \text{Attn}(\mathbf{s}_h) := \alpha \left(N_Q(Q) N_K(K)^\top\right) V = \alpha\big(\widetilde{Q}\widetilde{K}^\top\big) V \in \mathbb{R}^{M\times N}, \tag{9}$$

where $N_Q(\cdot)$ and $N_K(\cdot)$ 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, which takes inspiration from the normalization in the index map kernel integral (14) and (23), see also Boyd (2001) where the normalization for orthogonal bases essentially uses the (pseudo)inverse of the Gram matrices. In practice, its cheap alternative is chosen to be layer normalization (Ba et al., 2016) or batch normalization (Ioffe & Szegedy, 2015). $\alpha = h^2$ is a mesh-based weight such that the summation becomes an approximation to an integral quadrature.

To elaborate these rationales, the j-th column of the i-th row U_i of U and the i-th row $A_{i\bullet}$ are

$$(U_i)^j = \alpha\, A_{i\bullet} \cdot V^j, \quad \text{where} \quad A_{i\bullet} := (\widetilde{Q}\widetilde{K}^\top)_{i\bullet} \quad \text{and} \quad V^j := V_{\bullet j} \tag{10}$$

Thus, we can derive the dot-product attention in a basis expansion representation for U_i

$$U_i = \alpha\, A_{i\bullet} \begin{pmatrix} \text{——} & V_1 & \text{——} \\ \text{——} & \vdots & \text{——} \\ \text{——} & V_M & \text{——} \end{pmatrix} = \sum\nolimits_{m=1}^M \alpha\, A_{im} V_m =: \sum\nolimits_{m=1}^M \mathcal{A}(Q_i, K_m)\, V_m. \tag{11}$$

Here, $\alpha A_{i\bullet}$ contains the coefficients for the linear combination of $\{V_m\}_{m=1}^M$. This set $\{V_m\}_{m=1}^M$ form the V's row space, and it further forms each row of the output U by multiplying with A. $\mathcal{A}(\cdot,\cdot)$ in (11) stands for the attention kernel, which aggregates the pixel-wise feature maps to measure how the input 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 layer-wise.

If we further assume that there exist a set of feature maps for query, key, and value, e.g., see Choromanski et al. (2021). For U, the feature map $u(\cdot)$ that maps $\mathbb{R}^2 \to \mathbb{R}^{1\times N}$, i.e., $U_i = u(z_i) = [u_1(z_i), \cdots, u_N(z_i)]$, then a kernel $\kappa(\cdot,\cdot) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ can be defined as follows,

$$\mathcal{A}(Q_i, K_j) = \alpha\langle Q_i, K_j\rangle = \alpha\langle q(z_i), k(z_j)\rangle =: \alpha\,\kappa(z_i, z_j). \tag{12}$$

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

$$u_l(z) = \alpha \sum_{x\in\mathcal{M}} \big(q(z)\cdot k(x)\big) v_l(x)\,\delta_x \approx \int_\Omega \kappa(z,x) v_l(x)\,d\mu(x), \quad 1 \le l \le N. \tag{13}$$

After plugging $u(\cdot)$ in certain minimization such as (8), the backpropagation updates weights, which further lead 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 (13).

To connect attention with inverse problems, the multiplicative structure in a kernel integral form for attention (13) is particularly useful. This falls into the category of Pincherle-Goursat (degenerate) kernels (Kress, 1999, Chapter 11), and its approximability depends on the number of the expanded channels N. N decides number of learned basis functions in expansion (11), a subset of which eventually is used to form a set of basis. Here we show the following theorem, heuristically it says that: given enough but finite channels of latent representations, the attention kernel integral is able to "bootstrap" in frequency domain, that is, generating an output representation with higher frequencies than the input. Similar approximation result is impossible for CNN if opting for the usual framelet/wavelet interpretation (Ye et al., 2018). The full proof with a more rigorous setting is in Appendix D.

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 (e.g., Fourier-type $\sin(xz)$) to an error of $O(\epsilon)$ under certain norm $\|\cdot\|_X$. Then, there exists a set of $\{W^Q, W^K, W^V\}$ such that certain channel k' in the output of (11) 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 high-frequency eigenpairs of Λ_σ based on the low-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 result in Theorem 2 justifies the advantages of the attention mechanism for the considered problem.

3.3 From harmonic extension to tensor-to-tensor

To establish the connection between the problem of interest with the attention used in the Transformers, we begin with 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 $\mathcal{I}^D$ inspired by 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 a larger family of inverse problems. For instance, an approximation modified from the one in Chow et al. (2014) reads as

$$\mathcal{I}^D(x) \approx \mathcal{I}_1^D(x) := R(x)\,(\mathbf{d}(x)\cdot\nabla\phi(x)) \quad x \in \Omega,\ \mathbf{d}(x) \in \mathbb{R}^2, \tag{14}$$

where $R(x) = \big(\|f - \Lambda_{\sigma_0} g\|_{L^2(\partial\Omega)} |\eta_x|_Y\big)^{-1}$, and $f - \Lambda_{\sigma_0} g = (\Lambda_\sigma - \Lambda_{\sigma_0})g$ is the boundary data measuring the difference of NtD mappings. In particular, ϕ is the solution to (7) with $l = L = 1$ but with certain noise ξ on the boundary, i.e., $f - \Lambda_{\sigma_0} g$ is replaced by $f - \Lambda_{\sigma_0} g + \xi$, which is called the *harmonic extension*. The vector function $\mathbf{d}(x) = [d_1(x), d_2(x)]$ is called probing direction and can be chosen empirically as $\mathbf{d}(x) = \nabla\phi(x)/\|\nabla\phi(x)\|$. Lastly, η_x is the solutions to:

$$-\Delta\eta_x = -\mathbf{d}(x)\cdot\nabla\delta_x \quad \text{in} \quad \Omega, \quad \mathbf{n}\cdot\nabla\eta_x = 0 \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega} \eta_x\,\mathrm{d}s = 0, \tag{15}$$

where $\{\eta_x\}$ restricted on boundary $\partial\Omega$ is equipped with $|\cdot|_Y$ semi-norm, and δ_x is δ function associated with the point x. 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. Essentially, (14) suggests that the derivative of ϕ along a direction $\mathbf{d}(x)$ scaled by a quantity $R(x)$ approximates $\mathcal{I}^D(x)$, where $R(x)$ and $\mathbf{d}(x)$ are based on empirical choices. Indeed, the reconstruction accuracy is much limited by a single measurement, the nonparametric ansatz, and empirical choices. These restrictions give room for DL methodology.

Despite being restrictive, the formulation in (14) offers inspiration to give a potential answer to (Q1): the *harmonic extension* ϕ of $(f - \Lambda_{\sigma_0} g)$ can be used as the input to a *tensor2tensor*-type DNN. Constructing harmonic extension (2D features) from boundary data (1D signal input with limited embedding dimensions) 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 enable the reconstruction being highly robust with respect to the noise (e.g., see Figure 1). Second, in terms of using certain backbone network 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.

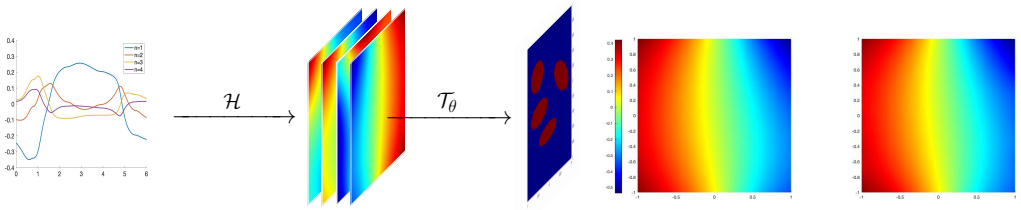

Figure 1: The left plot: a decomposition of the inverse operator into a harmonic extension operator $\mathcal{H}$ and a neural network $\mathcal{T}_\theta$: the 1D boundary data are preprocessed by extending to the domain interior to generate 2D feature maps which are then input for reconstruction. The right two plots: ϕ without and with 20% noise of ξ showing that harmonic extension is robust to the boundary noise.

The information of σ is deeply hidden in ϕ. As shown in Figure 1, one cannot observe any pattern of σ directly from ϕ, for more examples see Appendix B. It is different from and much more difficult 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.

As shown in Figure 1, $\mathcal{G}_\theta$ can be nicely decoupled into a composition of a parametrized neural network operator $\mathcal{T}_\theta$ and a non-learnable harmonic extension feature map $\mathcal{H}$, i.e., $\mathcal{G}_\theta = \mathcal{T}_\theta \circ \mathcal{H}$.

3.4 From index map integral to Transformer

In this last subsection, the probing direction $\mathbf{d}(x)$, the inner product $\mathbf{d}(x) \cdot \nabla\phi(x)$, and the norm $|\cdot|_Y$ in (14) are used as ingredients to form certain non-local 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 (9).

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 map in (14).

- The reformulation of the index function is motivated by the heuristics that the global information of ϕ should be used to locate a point x.

$$\hat{\mathcal{I}}_1^D(x) := R(x) \int_\Omega \mathbf{d}(x) \cdot \mathcal{K}(x, y) \nabla\phi(y) \, \mathrm{d}y. \tag{16}$$

  If $\mathcal{K}(x, y) = \delta_x(y)$, then (16) reverts to the original one in (14).
- Similarly, the probing direction $\mathbf{d}(x)$ is reasonably assumed to have a global dependence on ϕ

$$\mathbf{d}(x) := \int_\Omega \mathcal{Q}(x, y) \nabla\phi(y) \, \mathrm{d}y. \tag{17}$$

- In the quantigy $R(x)$ in (14), the key is $|\cdot|_Y$ which is assumed to have the following form:

$$|\eta_x|_Y^2 := (\mathcal{V}\eta_x, \eta_x)_{L^2(\partial\Omega)}. \tag{18}$$

  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 (14) can achieve maximum values for points inside D.

Based on the assumptions from (16) to (18), we derive a matrix representation approximating the new index function on a grid, which accord well with an attention-like architecture. Denote by

- $\boldsymbol{\phi}_n$: the vector that interpolates $\partial_{x_n}\phi$ at the grid points $\{z_j\}$, $n = 1, 2$.
- B: the discrete Laplacian on Ω_h of a finite element/finite difference discretization of $-\Delta$ coupled with the Neumann boundary condition and the zero integral normalization condition.
- R: the matrix that projects a vector defined at $\{z_i\}$ to one defined at the nodes on the boundary $\partial\Omega$.

We shall discretize the variable x by grid points z_i in (16) and obtain an approximation to the integral:

$$\int_\Omega \mathcal{K}(z_i, y)\partial_{x_n}\phi(y)\,\mathrm{d}y \approx \sum\nolimits_j \omega_j \mathcal{K}(z_i, z_j)\partial_{x_n}\phi(z_j) =: \mathbf{k}_i^T \boldsymbol{\phi}_n. \tag{19}$$

We then consider (17) and focus on one component $d_n(x)$ of $\mathbf{d}(x)$. With a suitable quadrature rule to compute the integral, it can be rewritten as

$$d_n(z_i) \approx \sum\nolimits_j \omega_j \mathcal{Q}(z_i, z_j)\partial_{x_n}\phi(z_j) =: \mathbf{q}_i^T \boldsymbol{\phi}_n. \tag{20}$$

Next, we proceed to express $\|\eta_{z_i}\|_Y$ by discretizing the variational form in (15) by a linear finite element method. Let $\{\psi^{(j)}\}_{j=1}^M$ be the set of the basis functions, and $\boldsymbol{\psi}_{n,i}$ be the vector approximating $[\partial_{x_n}\psi^{(j)}(z_i)]_{j=1}^M$ for each fixed z_i. Denote $\boldsymbol{\eta}_i$ as the vector approximating $\{\eta_{z_i}(z_j)\}_{j=1}^M$ and $\bar{\boldsymbol{\eta}}_i$ as its restriction on the boundary nodes. Then, the finite element discretization yields the linear systems:

$$B\boldsymbol{\eta}_i = \sum\nolimits_n d_n(z_i)\boldsymbol{\psi}_{n,i} \approx \sum\nolimits_n \boldsymbol{\psi}_{n,i}\mathbf{q}_i^T\boldsymbol{\phi}_n, \quad \text{and} \quad \bar{\boldsymbol{\eta}}_i := \eta_{z_i}|_{\partial\Omega} \approx \sum\nolimits_n RB^{-1}\boldsymbol{\psi}_{n,i}\mathbf{q}_i^T\boldsymbol{\phi}_n. \tag{21}$$

Note that the self-adjoint positive definite operator $\mathcal{V}$ can be parameterized by a symmetric positive definite (SPD) matrix denoted by V. We can approximate $|\eta_{z_i}|_Y^2$ as

$$|\eta_{z_i}|_Y^2 \approx \bar{\boldsymbol{\eta}}_i^T V \bar{\boldsymbol{\eta}}_i \approx \sum\nolimits_n c_{n,i}\boldsymbol{\phi}_n^T\mathbf{q}_i\mathbf{q}_i^T\boldsymbol{\phi}_n \tag{22}$$

where $c_{n,i} = \boldsymbol{\psi}_{n,i}^T B^{-1}R^T V R B^{-1}\boldsymbol{\psi}_{n,i} \geq 0$ as V is SPD. Define $\mathbf{v}_{n,i} = \sqrt{c_{n,i}}\mathbf{q}_i$ which can be considered as another learnable vector since $c_{n,i}$ comes from the learnable matrix V. Putting (19), (20) and (18) into (16), we obtain

$$\hat{\mathcal{I}}_1^D(z_i) \approx \Big\{\|f - \Lambda_{\sigma_0}g\|_{L^2(\partial\Omega)}^{-1}\big(\sum_n \boldsymbol{\phi}_n^T\mathbf{v}_{n,i}\mathbf{v}_{n,i}^T\boldsymbol{\phi}_n\big)^{-1/2}\Big\}\sum_n \boldsymbol{\phi}_n^T\mathbf{q}_i\mathbf{k}_i^T\boldsymbol{\phi}_n. \tag{23}$$

Now, using the notation from Section 3.2, we denote the learnable kernel matrices and an input vector: for $\diamond \in \{Q, K, V\}$, and $\mathbf{u} \in \{\mathbf{q}, \mathbf{k}, \mathbf{v}\}$

$$W^\diamond = \begin{bmatrix} \mathbf{u}_{1,1} & \cdots & \mathbf{u}_{1,M} \\ \mathbf{u}_{2,1} & \cdots & \mathbf{u}_{2,M} \end{bmatrix} \in \mathbb{R}^{2M\times M}, \quad \mathbf{p}_h = [\ \boldsymbol{\phi}_1 \quad \boldsymbol{\phi}_2\ ] \in \mathbb{R}^{1\times 2M}. \tag{24}$$

Then, we are able to rewrite (23) as

$$[\hat{\mathcal{I}}_1^D(z_i)]_{i=1}^M \approx C_{f,g}(\mathbf{p}_h W^Q * \mathbf{p}_h W^K)/(\mathbf{p}_h W^V * \mathbf{p}_h W^V)^{1/2} \tag{25}$$

where $C_{f,g} = \|f - \Lambda_{\sigma_0}g\|_{L^2(\partial\Omega)}^{-1}$ is a constant, and both $*$ and $/$ are taken element-wise. Here, we may define $Q = \mathbf{p}_h W^Q$, $K = \mathbf{p}_h W^K$, and $V = \mathbf{p}_h W^V$ as the values, keys and query. We can see that, the right matrix multiplications in (9) are low-rank approximations to the ones above in the attention mechanism. Hence, based on (25), essentially we need to find a function $\mathbf{I}$ resulting in a vector approximation to the true characteristic function $\{\mathcal{I}^D(z_j)\}$

$$\mathbf{I}(Q, K, V) \approx [\mathcal{I}^D(z_i)]_{i=1}^M. \tag{26}$$

Thus, the expression in (25) and (26) reveal that a Transformer is able to capture the classical formula in (14) equipped with non-local learnable kernels. Moreover, when there are L data pairs, the data functions ϕ_l are generated by computing their harmonic extensions as in (7). Then, each harmonic extension is then treated as a channel of the input $\mathbf{p}_h$.

The derivation above motivates that the attention mechanism nicely fits the underlying mathematical structure. In summary, we propose to use a Transformer-based deep direct sampling method, and subsequently show that it has significantly better performance than the widely used CNN-based U-Net in linear inverse problems and modern operator learner such as Fourier Neural Operator. In this regard, we provide a potential answer to the question (Q2); namely, the attention-based Transformer is better suited as it conforms more with the underlying mathematical structure, and both enjoy a global kernel formulation using the input data, which match 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 using a *single* channel ($L = 1$) of the 2D harmonic extension feature from the 1D signal input. There are two baseline models to compare: one is the CNN-based U-Net (Ronneberger et al., 2015), the other is the state-of-the-art operator learner Fourier Neural Operator (FNO) (Li 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 (9), 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). For more details of the data setup, training, evaluation in the all experiments please refer to Appendix B.

The comparison result can be found in Table 1. Due to the fact that FNO keeping only the modes in the lower end of the spectra, it performs relatively poor in this EIT benchmark that needs to recover traits that consist of higher modes (sharp boundary edges of inclusion) from lower modes (smooth harmonic extension). Thanks to Theorem 2, attention-based models are capable to recover "high frequency target from low frequency data", and in general outperform the CNN-based U-Net despite having only $1/3$ of the parameters. Another highlight is that, thanks to the unique PDE-based feature map through harmonic extension, the proposed models are extremely robust to noise. 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 B 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
FNO2d big	0.386	0.482	0.501	0.310	0.465	0.499	0.638	0.601	0.580	33.6m
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 some 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: similar to other operator learners, the data manifold on which the operator is learned is assumed to exhibit certain low-dimensional attributes that can be reasonably well approximated by a fixed number of bases.

REPRODUCIBILITY STATEMENT

This paper is reproducible. Experimental details about all empirical results described in this paper are provided in Appendix B. Additionally, we provide the PyTorch (Paszke et al., 2019) code for reproducing in the supplemental material. The dataset used in this paper is available at https://www.kaggle.com/datasets/anonymousauthor25/eit-transformer. Formal proofs under a rigorous setting of all our theoretical results are provided in Appendices C-D.

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. doi: https://doi.org/10.1016/j.camwa.2020.10.009. 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.

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

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. doi: 10.1155/2007/83016. 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.

George Barbastathis, Aydogan Ozcan, and Guohai Situ. On the use of deep learning for computational imaging. *Optica*, 6(8):921–943, Aug 2019. doi: 10.1364/OPTICA.6.000921. URL http://www.osapublishing.org/optica/abstract.cfm?URI=optica-6-8-921.

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.

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.

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

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

Shuhao Cao. Choose a transformer: Fourier or Galerkin. *arXiv preprint arXiv:2105.14995*, 2021.

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. doi: 10.1088/0266-5611/17/4/301. 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. doi: https://doi.org/10.1118/1.1534109. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1534109.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR 2021)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Matthew M. Dunlop and Andrew M. Stuart. The bayesian formulation of eit: Analysis and algorithms. *arXiv:1508.04106v2*, 2015.

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 non-standard wavelet form. *J. Comput. Phys.*, 384:1–15, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2019.02.002. 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. doi: 10.1117/1.JBO.24.5.051407. 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. doi: 10.3390/electronics7120422. URL https://www.mdpi.com/2079-9292/7/12/422.

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.

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.

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

R. Guo and J. 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, 2018.

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.

Martin Hanke and Martin Brühl. Recent progress in electrical impedance tomography. 19(6):S65–S90, nov 2003. doi: 10.1088/0266-5611/19/6/055. 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. doi: 10.1088/1361-6420/ab08be. URL https://doi.org/10.1088/1361-6420/ab08be.

A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. 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. doi: 10.1109/WACV51458.2022.00181. 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. doi: 10.1109/TMI.2021.3108910.

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. doi: https://doi.org/10.1016/j.ifacol.2018.07.114. URL https://www.sciencedirect.com/science/article/pii/S2405896318308589. 15th IFAC Conference on Programmable Devices and Embedded Systems PDeS 2018.

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.

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.

Kazufumi Ito, Bangti Jin, and Jun Zou. A direct sampling method for inverse electromagnetic medium scattering. *Inverse Probl.*, 29(9):095018, sep 2013. doi: 10.1088/0266-5611/29/9/095018. 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. doi: https://doi.org/10.1016/j.jcpx.2019.100003. URL https://www.sciencedirect.com/science/article/pii/S2590055219300022.

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

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*, 2022.

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). doi: 10.1002/mp.12344.

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. doi: 10.1088/1361-6420/aacf1d. 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. doi: 10.1137/060656930. 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.

Rainer Kress. *Linear Integral Equations*. Springer New York, 1999. doi: 10.1007/978-1-4612-0559-3. 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. doi: 10.1109/ICOMET.2019.8673502.

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. doi: 10.1109/ISBI.2011.5872603.

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

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*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

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. doi: 10.1038/s42256-021-00302-5. URL https://doi.org/10.1038/s42256-021-00302-5.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*, 2021.

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.

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.

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.

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. doi: 10.1117/1.OE.57.4.043111. 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.

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. doi: 10.1007/978-3-030-87589-3_28. URL https://doi.org/10.1007/978-3-030-87589-3_28.

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. doi: 10.1109/TIM.2019.2954722.

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

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. doi: 10.1109/TMI.2021.3077484.

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. doi: https://doi.org/10.1016/j.injury.2022.04.013. 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. doi: https://doi.org/10.1016/j.jqsrt.2008.08.006. URL https://www.sciencedirect.com/science/article/pii/S0022407308001854.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations (ICLR 2021)*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

J. Nasehi Tehrani, A. McEwan, C. 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. doi: https://doi.org/10.1016/j.apm.2011.07.055. 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. doi: 10.18653/v1/D19-1443. 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.

P.J. Vauhkonen, M. Vauhkonen, T. Savolainen, and J.P. Kaipio. Three-dimensional electrical impedance tomography based on the complete electrode model. *IEEE Trans. Biomedical Engrg.*, 46(9):1150–1160, 1999. doi: 10.1109/10.784147.

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. doi: 10.1109/JSEN.2020.3019309.

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. doi: 10.1063/1.4760253. 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.

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.

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). doi: 10.1038/nature25988.

A TABLE OF COMMON-USED NOTATIONS

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

Notation	Meaning
$\mathbb{P}$, $\mathbb{M}$	admissible Sobolev spaces for the inverse problems, which are $\mathbb{P} = L^2(\Omega)$ and $\mathbb{M} = \left[H^{-1/2}(\partial\Omega) \times H^{1/2}(\partial\Omega)\right]^L$ for EIT with L data pairs
∇u	the gradient vector of a function, $\nabla u(x) = (\partial_{x_1} u(x), \partial_{x_2} u(x))$
δ_x	the delta function such that $\int_\Omega f(y)\delta_x(y)\,\mathrm{d}y = f(x), \forall f$.
$\nabla u \cdot \mathbf{n}$	normal derivative of u, measures the rate of change along the direction of $\mathbf{n}$
Λ_σ	NtD mapping 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)
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
$H^s(\Omega)$, $s \geq 0$	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 EXPERIMENT SET-UP

B.1 DATA GENERATION AND TRAINING

In the numerical examples, the data generation mainly follows what were featured in (Fan & Ying, 2020; Guo & Jiang, 2020). For an example 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 generated by four random ellipses with the length of the semi-major axis and semi-minor axis sampled from $\mathcal{U}(0.1, 0.2)$ and $U(0.2, 0.4)$, respectively. The rotation angle sampled from $\mathcal{U}(0, 2\pi)$. There are 10800 samples in the training set, and 2000 in the testing set. The batch size is set to be 10. All models are trained using the `1cycle` scheduler (Smith & Topin, 2019) for 50 epochs with a 10-epoch warm-up.

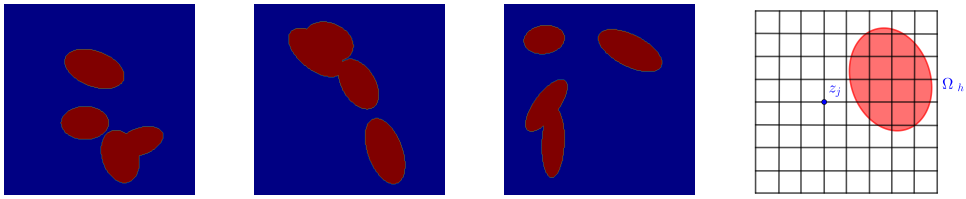

Figure 2: Randomly selected samples of elliptic inclusion to represent the coefficient σ that are used for training (left 1-3). A Cartesian mesh Ω_h with a grid point $\{z_j\}$ (right). In computation, discretization of $\mathcal{I}^D$ consists of values taken as 1 at the mesh points of Ω_h inside D and 0 at others.

The noise $\xi = \xi(x)$ below (14) is assumed to be

$$\xi(x) = (f(x) - \Lambda_{\sigma_0} g(x))\tau G(x) \tag{27}$$

where τ specifies the percentage 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 though the true data is smooth.

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

$$\mathcal{L}(\mathbf{p}_h, \mathbf{u}_h) := -\sum\nolimits_{z \in \mathcal{M}} \big(\mathbf{p}_h(z) \ln(\mathbf{u}_h(z)) + (1 - \mathbf{p}_h(z)) \ln(1 - \mathbf{u}_h(z))\big). \tag{28}$$

Thanks for 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 a total number of 50 epochs with no extra regularization such as weight decay. 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. Some testing results can be seen in Figure All models are trained on an RTX 3090 or an A4000. The codes to replicate the experiments is open-source and publicly available [1] .

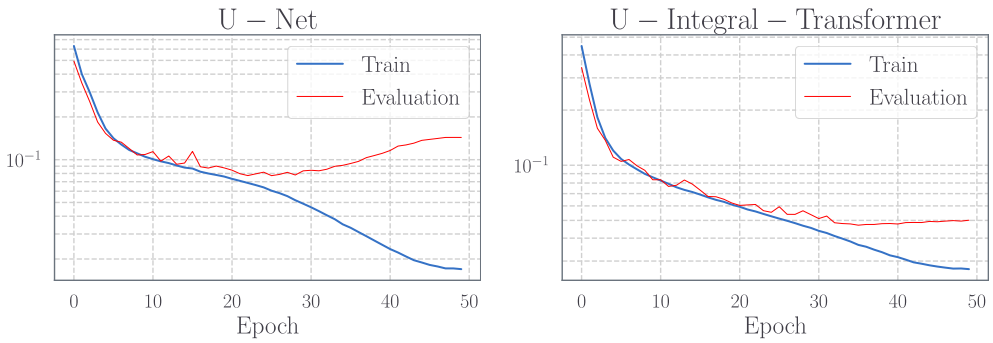

Figure 3: The left: the training-testing pixel-wise binary cross entropy convergence for the CNN-based U-Net with 31 million parameters, a clear overfitting pattern is shown. The right: the training-testing convergence for the attention-based U-Transformer with 11.4 million parameters.

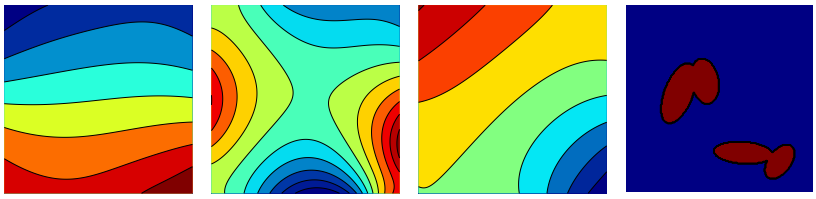

Figure 4: The harmonic extension feature map ϕ_l (left 1-3 as different channels' inputs to the neural network) corresponding to a randomly chosen sample's inclusion map (right). No visible relevance shown with the ground truth.

B.2 Network Architecture

The U-Transformer architecture is a drop-in replacement of the standard CNN-based U-Net. For a high level encoder-decoder schematics please see Figure 7.

Positional embedding. At each resolution, The 2D Euclidean coordinates of a $\sqrt{M} \times \sqrt{M}$ regular grid are applied a channel expansion through a learnable linear layer, and then are added to each latent representation.

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) shall be replaced by the layer normalization (Ba et al., 2016).

[1] Please download the code in the supplemental material zip file and follow the instructions in `README.md`.

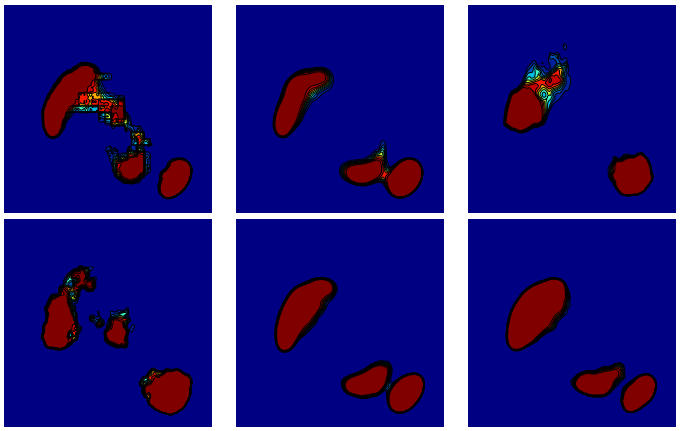

Figure 5: The neural network evaluation result for the inclusion in Figure 4 using various models. Top from left to right: U-Net baseline (7.7m) prediction with a single channel; U-Net big (31m) prediction with a single channel; Fourier Neural Operator (10.4m) prediction with 1 channel. Bottom from left to right: Fourier Neural Operator big (33m) prediction with 1 channel; UIT (11.4m) prediction with a single channel; UIT (11.4m) prediction with 3 channels.

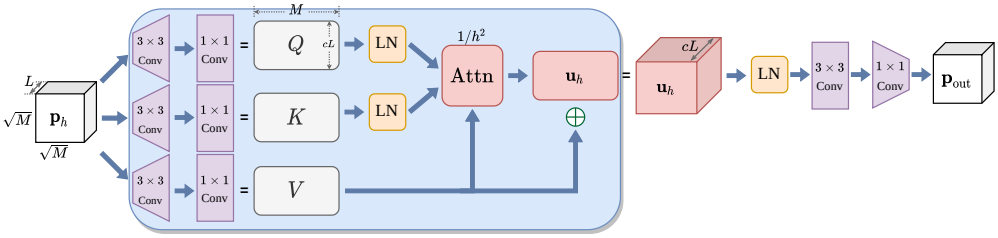

Figure 6: The modified 2D attention-based encoder layer. c: the channel expansion constant.

Mesh-normalized attention. The scaled dot-product attention in the network is chosen to be the softmax-free integral kernel attention in (9) with a mesh-based normalization. For a diagram in a single attention head please refer to Figure 6.

Coarse-fine interaction in up blocks. The convolution layer on the coarsest level is replaced by the attention with pre-inner-product normalization in Section 3.2. 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, such that the attention kernel to measure the interaction between different channels is built from on the coarse grid. V is associated with a finer grid. The major differences of ours with the one in Petit et al. (2021) are that ours are inspired by the kernel integral for a PDE problem, thus the modified attention in our method has (1) no softmax normalization or (2) no Hadamard product-type skip connection.

C Proof of Theorem 1

Theorem 1 (A finite dimensional approximation of the index map)**.** *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*

$$-\Delta\phi_l = 0 \quad in \quad \Omega, \quad \mathbf{n}\cdot\nabla\phi_l = (f_l - \Lambda_{\sigma_0} g_l) = (\Lambda_\sigma - \Lambda_{\sigma_0})g_l \quad on \quad \partial\Omega, \quad \int_{\partial\Omega} \phi_l \, \mathrm{d}s = 0, \tag{29}$$

where $l = 1, 2, \cdots$. Define the space on $\mathbb{R}^N$ (the spatial dimension $N = 2, 3$):

$$\mathbb{S}_L = \text{Span}\{\partial_{x_n}\phi_l \, \partial_{x_m}\phi_l : \; l = 1, ..., L, \; m, n = 1, \ldots, N\}. \tag{30}$$

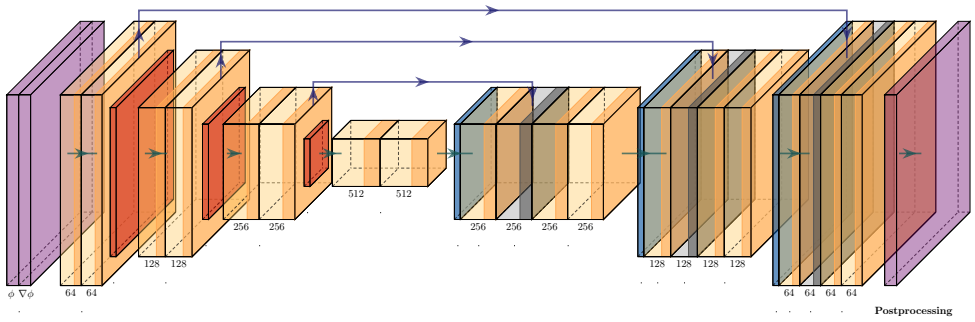

Figure 7: The schematics of the U-Transformer that follows the standard U-Net. The input is the concatenation of discretizations of ϕ and $\nabla\phi$. The output is the approximation to the index map $\mathcal{I}^D$. The numbers of latent basis functions (channels) are annotated below each representation. ■: 3×3 convolution + ReLU; ■: layer normalization or batch normalization; ■: bilinear interpolation from the fine grid to the coarse grid; ■: cross attention from the coarse grid to the fine grid; ■: input and output discretized functions in certain Hilbert spaces.

Then, for any $\epsilon > 0$, there exists a sufficiently large L such that

$$\inf_{\mathcal{I}_L^D \in \mathbb{S}_L} \|\mathcal{I}^D - \mathcal{I}_L^D\|_{L^\infty(\Omega)} \leq \epsilon. \tag{31}$$

Proof. Let $\mathbf{d}$ be any probing direction in $\mathbb{R}^N$ (Brühl, 2001; Hanke & Brühl, 2003). Define a function

$$\Theta_L(x) = \sum_{l=1}^{L} \frac{(\mathbf{d} \cdot \nabla\phi_l(x))^2}{\lambda_l^3}. \tag{32}$$

By Theorem 4.1 in (Guo & Jiang, 2020), we can show that

$$\lim_{L\to\infty} \Theta_L(x) = \begin{cases} \infty, & \text{if } x \notin D, \\ \text{a finite constant}, & \text{if } x \in D. \end{cases} \tag{33}$$

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 interger L such that $\Theta_L(x) > 4\rho\epsilon^{-2}/\pi^2$, $\forall x \notin D$. Define

$$\mathcal{I}_L^D(x) = 1 - \frac{2}{\pi}\arctan\left(\frac{\pi\epsilon}{2\rho}\Theta_L(x)\right) \tag{34}$$

Note the fundamental inequality $z > \arctan(z) \geq \frac{\pi}{2} - z^{-1}$, $\forall z > 0$. 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

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

Therefore, the function in (34) fulfills (31).

□

D Proof of Theorem 2

In presenting Theorem 2 in Section 3.2, we use the term "multiplicative" to describe the fact that two latent representations are multiplied in the attention mechanism, and, in contrast, there is no such operation 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. If it is assumed that this discretization can represent functions of such a frequency with certain approximation error, then the resulting matrix/tensor can be an approximation of a function with higher frequency than the existing latent representation under the same discretization.

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 $N \geq 1$ is 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 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 (13), (iii) the discretization satisfies that $|\sum_{i=1}^M hf(z_i) - \int_\Omega f(x)\,\mathrm{d}x| \leq C\,h$. Then, for $v(x) = \sin(ax)$ for some $a \in \mathbb{R}^+$, there exists a set of attention weights $\{W^Q, W^K, W^V\}$ such that for $u(x) = \sin(a'x)$ with $a' > a$*

$$\|\tilde{u} - u\|_{L^2(\Omega)} \leq C \max\{h, \|\varepsilon\|_{L^\infty(\Omega)}\}, \tag{35}$$

where $\tilde{u}$ and $\tilde{\kappa}(\cdot,\cdot)$ are defined as the output of and the kernel of the attention formulation in (13), 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 $a = 1$ and $a' = 2$. 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

$$\kappa(x, z) := \sin(2(x - z)), \tag{36}$$

it is straightforward to verify that for $v(x) := \sin(x)$,

$$\int_\Omega \sin(2(x - z)) \sin(x)\,\mathrm{d}x = c_1 \sin(2z) =: c_1 u(z). \tag{37}$$

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

$$\kappa_N(x, z) := \sum_{l=1}^N \frac{(z - z_0)^{l-1}}{(l-1)!} \frac{\partial^{l-1}\kappa}{\partial z^{l-1}}(x, z_0). \tag{38}$$

It is straightforward to check that

$$\|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} \leq c_1 \frac{\sqrt{(\pi - z_0)^{2N+1} + z_0^{2N+1}}}{N!\sqrt{2N+1}} \left\| \frac{\partial^N \kappa}{\partial z^N}(x,\cdot) \right\|_{L^\infty(\Omega)}. \tag{39}$$

By the assumptions on $\kappa(\cdot,\cdot)$, and a simple calculus we have

$$\|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} \leq c_2 \frac{(\pi)^N}{N!\sqrt{N}} \left\| \frac{\partial^N \kappa}{\partial z^N}(x,\cdot) \right\|_{L^\infty(\Omega)}. \tag{40}$$

Next, denote $q_l(z) := (z - z_0)^{l-1}/(l-1)!$ and $k_l(x) := \partial^{l-1}\kappa/\partial z^{l-1}(x, z_0)$ for $1 \leq l \leq N$, i.e.,

$$\kappa_N(x, z) = \sum_{l=1}^N q_l(x) k_l(z). \tag{41}$$

By the choice of the latent representation space, $q_l \in Y = \mathrm{span}\{p_j\}$, thus there exists a set of weights $\{w_l^Q \in \mathbb{R}^L\}_{l=1}^N$ such that $q_l(\cdot) = \sum_{j=1}^l w_{l,j}^Q p_j(\cdot) =: \tilde{q}_l(\cdot)$. 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 estimate, there exists a set of weights $\{w_l^K \in \mathbb{R}^N\}_{l=1}^N$, such that $k_l(\cdot) \approx \sum_{j=1}^N w_{l,j}^K p_j(\cdot) =: \tilde{k}_l(\cdot)$, 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 $k_l(\cdot)$

$$\|k_l(\cdot) - \tilde{k}_l(\cdot)\|_{L^2(\Omega)} \leq c_3 \frac{\pi^N}{2^N (N+1)^N} |k_l(\cdot)|_{H^N(\Omega)}. \tag{42}$$

Similarly, for $v(\cdot) := v_1(\cdot)$ which can be chosen as the 1st channel of V defined earlier, we have $\{w_l^V \in \mathbb{R}^N\}$ such that $v(\cdot) \approx \sum_{j=1}^N w_{1,j}^K p_j(\cdot) =: \tilde{v}(\cdot)$

$$\|v(\cdot) - \tilde{v}(\cdot)\|_{L^2(\Omega)} \leq c_4 \frac{\pi^N}{2^N (N+1)^N} |v(\cdot)|_{H^N(\Omega)}. \tag{43}$$

Now, to approximate the frequency bootstrapping in (37), define

$$\tilde{u}(z) := \int_\Omega \tilde{\kappa}_N(x,z)\tilde{v}(x)\,\mathrm{d}x, \quad \text{with } \tilde{\kappa}_N(x,z) := \sum_{l=1}^N \tilde{q}_l(x)\tilde{k}_l(z). \tag{44}$$

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

$$\begin{aligned} u(z) - \tilde{u}(z) &= \int_\Omega \kappa(x,z)v(x)\,\mathrm{d}x - \int_\Omega \kappa_N(x,z)\tilde{v}(x)\,\mathrm{d}x \\ &= \int_\Omega \big(\kappa(x,z) - \tilde{\kappa}_N(x,z)\big)v(x)\,\mathrm{d}x + \int_\Omega \kappa_N(x,z)\big(v(x) - \tilde{v}(x)\big)\,\mathrm{d}x. \end{aligned} \tag{45}$$

Thus we have

$$\begin{aligned} \|u - \tilde{u}\|_{L^\infty(\Omega)} &\leq \max_{z\in\Omega} \left\{ \int_\Omega \left|\big(\kappa(x,z) - \tilde{\kappa}_N(x,z)\big)v(x)\right| \,\mathrm{d}x + \int_\Omega \left|\kappa_N(x,z)\big(v(x) - \tilde{v}(x)\big)\right| \,\mathrm{d}x \right\} \\ &\leq \max_{z\in\Omega} \Big\{ \underbrace{\|\kappa(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)}}_{(*)} \|v\|_{L^2(\Omega)} + \|\tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)} \|v - \tilde{v}\|_{L^2(\Omega)} \Big\}. \end{aligned} \tag{46}$$

Now, by triangle inequality, $q_l = \tilde{q}_l$, the definitions above (36) implying $|k_l(\cdot)|_{H^N(\Omega)} \leq c\,2^N$ and $\|q_l\|_{L^2(\Omega)} \leq c\pi^N/(\sqrt{N}N!)$, and the estimate in (42)

$$\begin{aligned} (*) &\leq \|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} + \|\kappa_N(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)} \\ &\leq \|\kappa(x,\cdot) - \kappa_N(x,\cdot)\|_{L^2(\Omega)} + \sum_{l=1}^N \|q_l\|_{L^2(\Omega)} \|k_l - \tilde{k}_l\|_{L^2(\Omega)} \\ &\leq c_5 \left(\frac{(2\pi)^N}{N!\sqrt{N}} + c_6 \frac{\pi^N}{2^N(N+1)^N} \sum_{l=1}^N \|q_l\|_{L^2(\Omega)} |k_l(\cdot)|_{H^N(\Omega)} \right) \\ &\leq c_6 \frac{(2\pi)^N}{N!\sqrt{N}}. \end{aligned} \tag{47}$$

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:

$$\begin{aligned} \|\tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)} &\leq \|\kappa_N(x,\cdot)\|_{L^2(\Omega)} + \|\kappa_N(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)} \\ &\leq \|\kappa(x,\cdot)\|_{L^2(\Omega)} + \|\kappa_N(x,\cdot) - \tilde{\kappa}_N(x,\cdot)\|_{L^2(\Omega)}, \end{aligned} \tag{48}$$

which can be further estimate reusing the argument in (47).

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

$$\|u - \tilde{u}\|_{L^2(\Omega)}^2 \leq \|u - \tilde{u}\|_{L^1(\Omega)} \|u - \tilde{u}\|_{L^\infty(\Omega)} \leq 2\max|u|\,\|u - \tilde{u}\|_{L^\infty(\Omega)}^2. \tag{49}$$

□