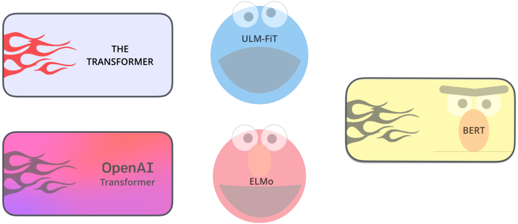
Transformer Meets Boundary Value Inverse Problems

Ruchi Guo (UC Irvine), Shuhao Cao (UMKC), and Long Chen (UC Irvine)

The Eleventh International Conference on Learning Representations (ICLR 2023)
arXiv:cs.LG/2209.14977
https://github.com/scaomath/eit-transformer

UMKC
School of Science and Engineering

Attention is all we need?



(left) Google search results on "Transformers". Credit: Paramount Pictures.

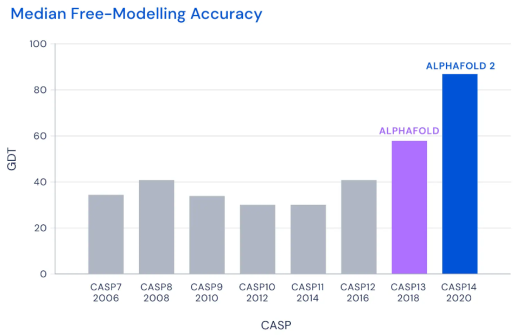
(Right) Jay Alammar: *The Illustrated Transformers: BERT, ELMo, and companies*.

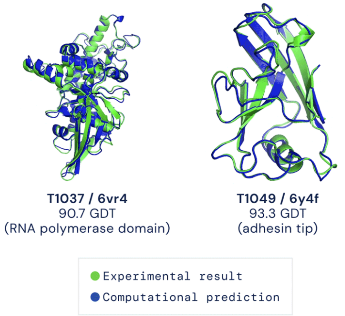
"Google replaced the original Google Neural Machine Translation system in Google Translate with a Transformer encoder[1] and an RNN decoder, implemented in `TensorFlow`." [2]

[1] A. Vaswani et al. (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems (NIPS 2017)*. Vol. 30.

[2] `https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html`

Attention is all we need?





July 2021: AlphaFold 2 uses a Transformer to map the input of a multiple sequence alignment (MSA) consisting amino acids to the output of the 3D structure of a protein.

Source: Nature & Deepmind. [3] [4]

[3] AlphaFold: a solution to a 50-year-old grand challenge in biology https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

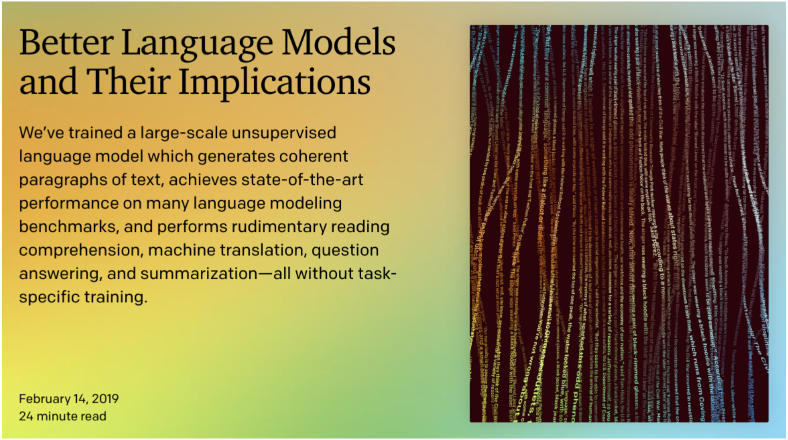
[4] J. Jumper et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589.

Attention is all we need?

(Left) Stable Diffusion by Stability AI[5]. (Right) AlphaTensor by Deepmind.

[5] R. Rombach et al. (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on CVPR*, pp. 10684–10695.

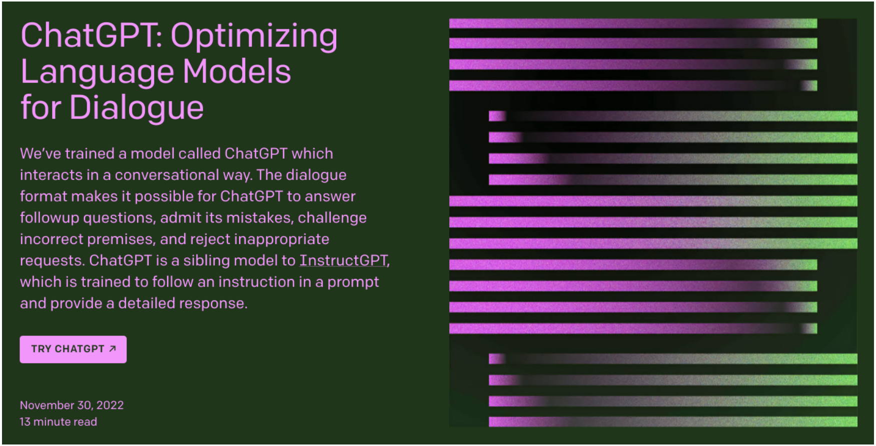
Attention is all we need?



GPT (117m), GPT-2 (1.2b), GPT-3 (175b). OpenAI[6].

[6] J. Kaplan et al. (2020). "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361*.

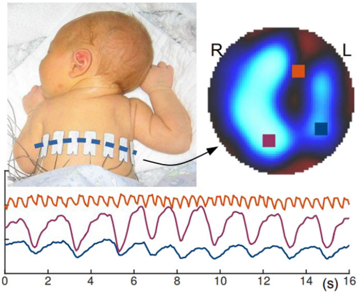
Attention is all we need?

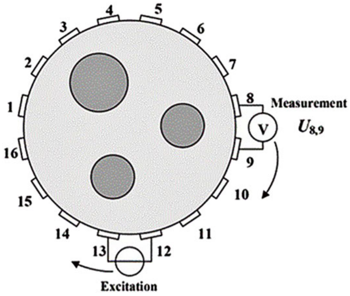


ChatGPT. OpenAI[7].

[7] N. Stiennon et al. (2020). "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems (NeurIPS 2020)* 33, pp. 3008–3021.

Electrical Impedance Tomography (EIT)



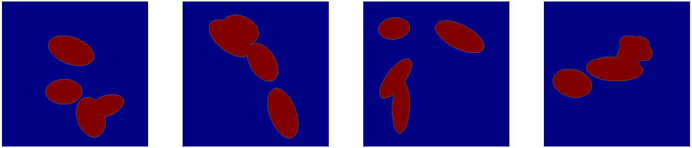


(Left)A 10-day-old infant with EIT electrodes[8]. By performing lung function imaging of newborns, timely diagnosis and treatment of lung diseases in early development of newborns without radiation damage can be done. (Right) Working principle of a 16 electrode system. Adjacent excitation is to select a pair of adjacent electrodes to input safe current, and then measure the output voltage between several pairs of adjacent electrodes except the excitation source.

[8] Y. Shi et al. (2021). "The research progress of electrical impedance tomography for lung monitoring". In: *Frontiers in Bioengineering and Biotechnology* 9.

How does “Transformer”, the backbone of all language models, has anything to do with “Boundary Value Inverse Problems” such as EIT?

The mathematical formulation of EIT



Several instances of inclusion $\sigma(\cdot)$.

The forward model of EIT

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

- Current: $g = \sigma \nabla u \cdot \boldsymbol{n}|_{\partial\Omega}$ (Neumann boundary condition)
- Voltages: $f = u|_{\partial\Omega}$ (Dirichlet boundary condition)

Neumann-to-Dirichlet (NtD) mapping:

$$\Lambda_\sigma : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega), \quad g = \sigma \nabla u \cdot \boldsymbol{n}|_{\partial\Omega} \xrightarrow{\text{solve } (\star)} f = u|_{\partial\Omega}.$$

Inverse Problem of EIT

Forward and inverse operator

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

- The measurement on $\partial\Omega$.
- The coefficient to be recovered.
- What we need (optimistically) is "knowing Λ_σ": for a set of basis $\{g_l\}_{l=1}^{\infty}$ of the corresponding Hilbert space, 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 $\boldsymbol{A}_\sigma$.

  $$\boldsymbol{f} = \boldsymbol{A}_\sigma \boldsymbol{g},$$

  where $\boldsymbol{g}$ and $\boldsymbol{f}$ are (infinite dimensional) vector representations of functions g and f.
- BCR-Net[9] is a DNN approximation of $\mathcal{F}^{-1}$ based on a large but finite sized matrix $\widetilde{\mathrm{A}}_\sigma$ as an accurate approximation to A_σ.

[9] Y. Fan and L. Ying (2020). "Solving electrical impedance tomography with deep learning". In: *Journal of Computational Physics* 404, p. 109119.

Inverse Problem of EIT

Question

What if the full spectrum of Λ_σ is not accessible?
Can we use only a few data pairs $\{(g_l, f_l)\}_{l=1}^{L}$ for reconstruction?

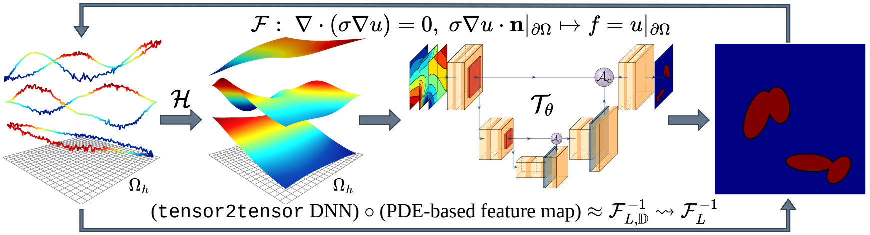
Forward and inverse operator with limited data pairs:

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

- Extremely ill-posed or even not well-defined: the same boundary measurements may correspond to different σ.[10].
- In $\boldsymbol{f} = \boldsymbol{A}_\sigma \boldsymbol{g}$, for $\boldsymbol{g}_l = \boldsymbol{e}_l$, $l = 1, \dots, L$, with $\boldsymbol{e}_l$ being unit vectors of a chosen basis, $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_L)$ only gives the first L columns of $\boldsymbol{A}_\sigma$.
- Restricting $\mathcal{F}_L^{-1}$ at a compact set of sampled data $\mathbb{D} := \{\sigma^{(k)}\}_{k=1}^{N}$.

[10] V. Isakov and J. Powell (1990). "On the inverse conductivity problem with one measurement". In: *Inverse Probl.* 6, p. 311.

From EIT to deep learning



- Learn an approximation to $\mathcal{F}^{-1}_{L,\mathbb{D}} : \{(g_1, \Lambda_{\sigma^{(k)}} g_1), \ldots, (g_L, \Lambda_{\sigma^{(k)}} g_L)\} \mapsto \sigma^{(k)}$.
- "Well-defined" enough as a high-dimensional interpolation (learning) problem on a compact data submanifold[11] with an end-to-end setting. Then generalization can be done for newly incoming σ's.
- The incomplete information of Λ_σ due to small L for one single σ is compensated by a large $N \gg 1$ sampling of different σ's.

[11] O. Ghattas and K. Willcox (2021). "Learning physics-based models from data: perspectives from inverse problems and model reduction". In: *Acta Numerica* 30, pp. 445–554.

From EIT to deep learning

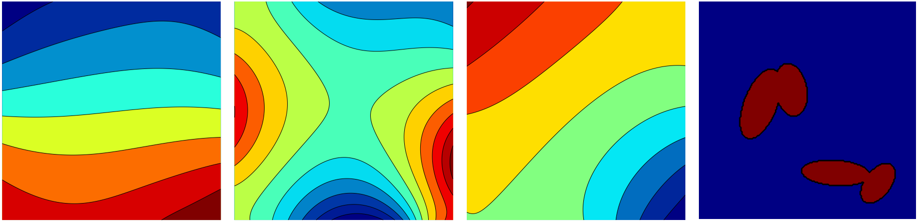
- What is an appropriate finite dimensional data format as inputs to the neural network?
- Is there a suitable neural network matching the mathematical structure?

Inspiration: direct sampling

Generate ϕ_l: the *harmonic extension* of $f_l - \Lambda_{\sigma_0} g_l$

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

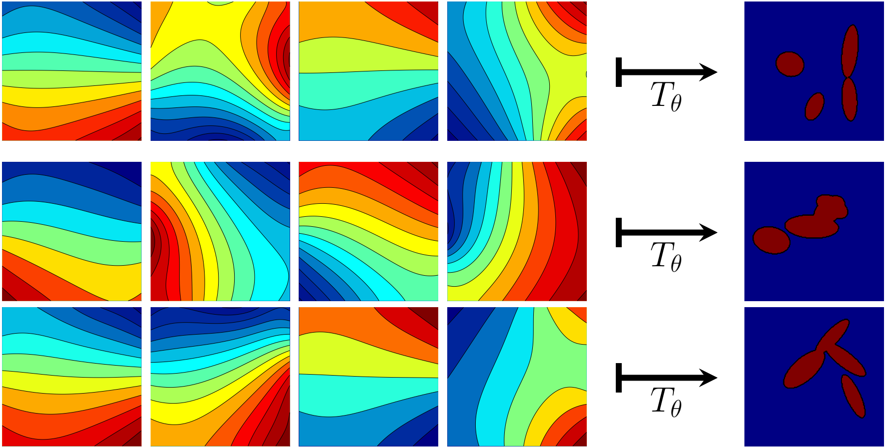
$$-\Delta \phi_l = 0 \quad \text{in} \quad \Omega, \quad \mathrm{n} \cdot \nabla \phi_l = (f_l - \Lambda_{\sigma_0} g_l) = (\Lambda_\sigma - \Lambda_{\sigma_0}) g_l \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega} \phi_l \mathrm{d}s = 0,$$



(a) (b) (c) (d)

EIT problem: (a)–(c) the input $\{\phi_l\}_{l=1}^L$ are harmonic extensions "features" for true σ (d).

Operator learning for EIT as a tensor2tensor problem



More examples of direct sampling: (Ideal) NtD map Λ_σ's whole spectrum ($L = \infty$) can recover the inclusion σ with various interfaces. (Practice) "learn" a *single* parametrized operator T_θ that maps (a few, $L \leq 3$) harmonic extension features to reconstruct the inclusions.

From positional embedding to the grid of a discretization

- Re-interpreting the latent representation in $\mathbb{R}^{n\times d}$ from:

$$\boxed{\text{Row} = \text{A word in a sentence}} \quad \text{to}$$

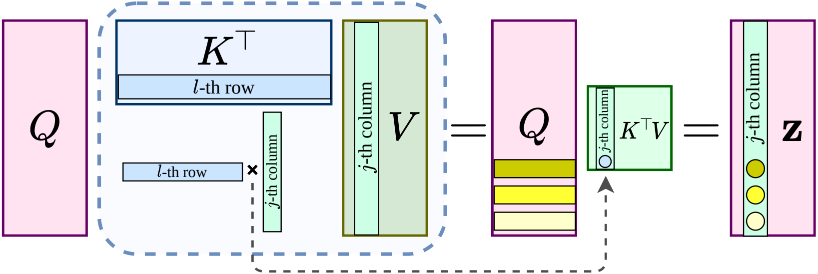
$$\boxed{\text{Column} = \text{A basis function in a Hilbert subspace.}}$$

$$\boxed{\begin{matrix}(1,0)\\(0,1)\end{matrix}} = \begin{matrix}\boxed{\text{Machine}}\\\boxed{\text{Learning}}\end{matrix} \subset \mathrm{Col}\left\{\binom{1}{0}, \binom{0}{1}\right\}$$

- The columns of Query/Key/Value contain the learned basis functions spanning certain subspaces of different Hilbert spaces.
- The column spaces of Query/Key/Value will be enriched by $\mathsf{span}\{w_j \in \mathbb{X}_h : w_j(x_i) = (\sigma_s(\mathrm{x}))_{ij}, 1 \le j \le d\} \subset \mathcal{H}$ to try to capture how an operator of interest responses to the subset of inputs.

Galerkin-type attention inspired by PDE

- In this linear attention regime: Q: values, K: query, V: keys.
- If K and V are orthornormal pairwise, then this is a (learnable) Petrov-Galerkin projection **!**[12]



$$(\boldsymbol{z}^j)_i = z_j(x_i) = h\sum_{l=1}^{d}(\boldsymbol{k}^l \cdot \boldsymbol{v}^j)(\boldsymbol{q}^l)_i \approx \sum_{l=1}^{d}\left(\int_\Omega k_l(\xi)v_j(\xi)\mathrm{d}\xi\right) q_l(x_i).$$

[12] S. Cao (2021). "Choose a Transformer: Fourier or Galerkin". In: *Advances in Neural Information Processing Systems (NeurIPS 2021)*. Vol. 34. eprint: 2105.14995 (cs.LG).

Inspiration: direct sampling

Direct sampling method for EIT[13]: $f - \Lambda_{\sigma_0} g \to \phi \to \boldsymbol{d} \to \eta_x$.

$$\mathcal{I}_1^D(x) := \frac{\boldsymbol{d}(x) \cdot \nabla\phi(x)}{\|f - \Lambda_{\sigma_0} g\|_{L^2(\partial\Omega)} |\eta_x|_{H^s(\partial\Omega)}}.$$

where

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

- The empirical formula of $\mathcal{I}^D(x)$ can be written as an integral with Gaussian-like density, that attains maximum values for $x \in D$.
- The accuracy is much limited by some empirical choices of quantities such as the probing direction $\boldsymbol{d}(x)$ and $s = 3/2$.
- This type of simple formula in direct sampling can be derived only for a single data pair.

[13] Y. T. Chow, K. Ito, and J. Zou (2014). "A direct sampling method for electrical impedance tomography". In: *Inverse Probl.* 30.9, p. 095003.

From direct sampling to attention integral

- The global information of ϕ used as "keys" to locate a point x to probe.

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

- The probing direction $\boldsymbol{d}(x)$ as "query" is assumed to depend globally on ϕ

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

  Choice of the probing direction in direct sampling[14]: If $\mathcal{Q}(x,y) = \delta_x(y)/\|\nabla\phi(x)\|$, then $\boldsymbol{d}(x) = \nabla\phi(x)/\|\nabla\phi(x)\|$.

- In $R(x)$, $|\cdot|_Y$ is assumed to be $|\eta_x|_Y^2 := (\mathcal{V}\eta_x, \eta_x)_{L^2(\partial\Omega)}$, where η_x is the potential using the probing as source. If $\mathcal{V}$ induces a Gaussian-like kernel which the attention kernel does induce[15], the index function can achieve maximum values for points inside D.

[14] M. Ikehata (2000). "Reconstruction of the support function for inclusion from boundary measurements". In: *Journal of Inverse and Ill-posed Problems* 8.4, pp. 367–378.

[15] H. Peng et al. (2021). "Random Feature Attention". In: *International Conference on Learning Representations*.

An architectural advantage of $(QK^\top)V$

Theorem (Frequency bootstrapping (simplified informal 1D version)[16])

Suppose there exists a channel l in the current latent representation such that $(V_i)_l = \sin(az_i)$ for some $a \in \mathbb{Z}^+$, the current finite-channel sum attention kernel approximates a "nice" kernel to an error of $O(\epsilon)$ with only "lower frequency" modes. Then, there exists a set of weights such that certain channel k' in the output of the attention layer approximates $\sin(a'z)$, $\mathbb{Z}^+ \ni a' > a$ with comparable error.

- Heuristics: multiplicative neural architecture can use data-driven basis functions to characterize operators.

$$u_l(z) = h^2 \sum_{x\in\mathcal{M}} \big(q(z)\cdot k(x)\big)\, v_l(x)\, \delta_x \approx \int_\Omega \kappa_\theta(z,x) v_l(x)\, d\mu(x).$$

- Proof: use the tools of Pincherle-Goursat (degenerate) kernels for $\kappa_\theta(z,x;v) = \sum_{l=1}^N q_l(x;v) k_l(z;v)$.

[16] R. Guo, S. Cao, and L. Chen (2023). "Transformer Meets Boundary Value Inverse Problems". In: *The Eleventh International Conference on Learning Representations (ICLR 2023)*

Attention is all we need?

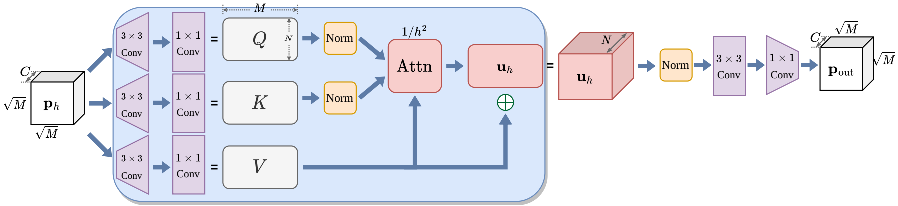


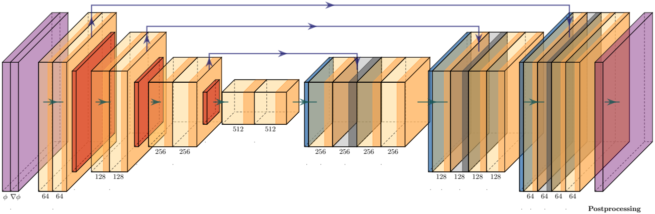
Figure: Schematics of a modified attention layer of the Transformer-based operator learner.

- Positional embedding: At each resolution, The 2D $\sqrt{M} \times \sqrt{M}$ Cartesian grid.
- ResNet **`DoubleConv`**: The double convolution block is modified[17] from that commonly seen in Computer Vision CNN[18].
- The "interaction" (attention matrix) between different latent representations can be computed using coarse latent representations.

[17] Z. Liu et al. (2022). "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986.

[18] K. He et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Drop-in replacement of U-Net



The schematics of the U-Transformer that follows U-Net[19][20]. 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; ■: normalization; ■: interpolation; ■: cross attention from the coarse grid to the fine grid; ■: input and output discretized functions.

[19] O. Ronneberger, P. Fischer, and T. Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

[20] R. Guo and J. Jiang (2021). "Construct Deep Neural Networks based on Direct Sampling Methods for Solving Electrical Impedance Tomography". In: *SIAM Journal on Scientific Computing* 43.3, B678–B711.

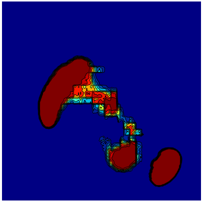
Electrical impedance tomography (EIT)

- Noise: $\xi = \xi(x)$ is assumed to be $\xi(x) = (f(x) - \Lambda_{\sigma_0} g(x))\tau G(x)$ where τ specifies the percentage of noise, and $G(x)$ is a Gaussian distribution.

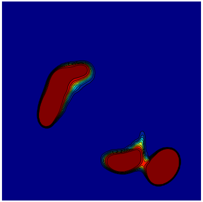
	Relative L^2 error			Position-wise cross entropy			Dice coefficient			# params
	$\tau=0$	$\tau=0.05$	$\tau=0.2$	$\tau=0$	$\tau=0.05$	$\tau=0.2$	$\tau=0$	$\tau=0.05$	$\tau=0.2$	
U-Net	0.200	0.341	0.366	0.0836	0.132	0.143	0.845	0.810	0.799	7.7m
FNO2d[21]	0.318	0.492	0.502	0.396	0.467	0.508	0.650	0.592	0.582	10.4m
Hybrid UT[22]	0.185	0.320	0.333	0.0785	0.112	0.116	0.877	0.829	0.821	11.9m
Cross-Attention UT[23]	0.171	0.305	0.311	0.0619	0.105	0.109	0.887	0.840	0.829	11.4m
U-Net+Coarse Attn	0.184	0.343	0.360	0.0801	0.136	0.147	0.852	0.807	0.804	8.4m
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

[21] Z. Li et al. (2021). "Fourier Neural Operator for Parametric Partial Differential Equations". In: *The Ninth International Conference on Learning Representations (ICLR 2021)*.
[22] Y. Gao, M. Zhou, and D. N. Metaxas (2021). "UTNet: a hybrid transformer architecture for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 61–71.
[23] H. Wang et al. (2022). "Mixed transformer u-net for medical image segmentation". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2390–2394.

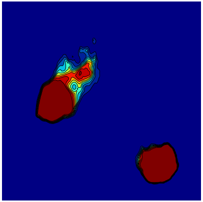
Reconstruction for unseen samples



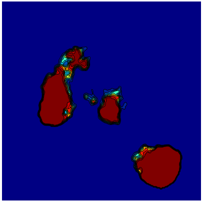
(a) U-Net (7.7m)
$L = 1$



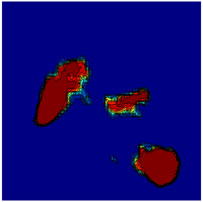
(b) U-Net (33m)
$L = 3$



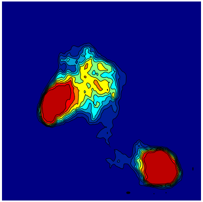
(c) FNO2d (10.4m)
$L = 1$



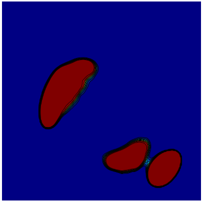
(d) FNO2d big (33m)
$L = 1$



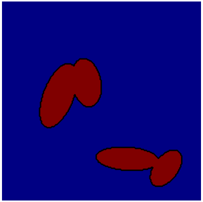
(e) Multiwavelet NO
(9.8m) $L = 1$



(f) Adaptive FNO
(10.7m) $L = 1$



(g) UIT (11.4m)
$L = 1$



(h) Ground truth
inclusion

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 in part by NSF grants DMS-2136075 and UMKC School of Science and Engineering computing facilities.
- The ICLR area chair for handling the paper, and anonymous reviewers for the suggestions to improve the paper.