CViT: Continuous Vision Transformer for Operator Learning

Sifan Wang, Jacob Seidman, Shyam Sankaran, Hanwen Wang, George J Pappas, Paris Perdikaris

Penn
University of Pennsylvania
ICLR
International Conference On Learning Representations
Yale

Operator Learning

Learning a function:

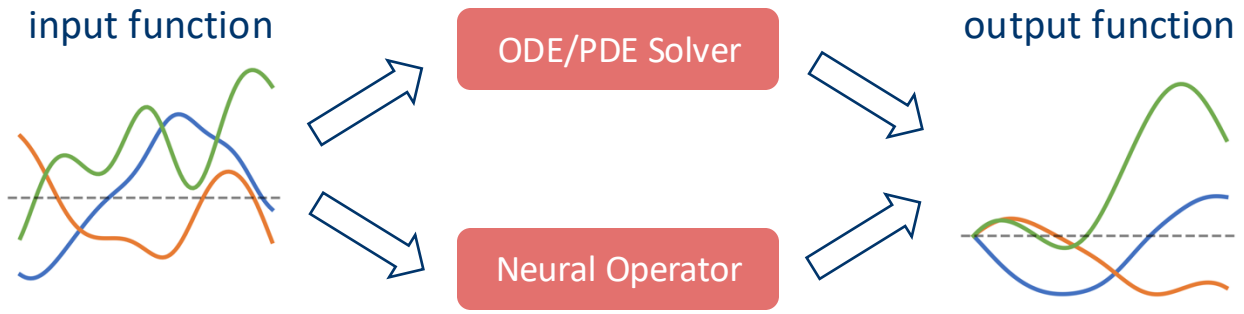
$$\mathbf{u}: \mathbb{R}^n \to \mathbb{R}^m$$
$$\mathbf{x} \to \mathbf{u}(x)$$

Learning an operator:

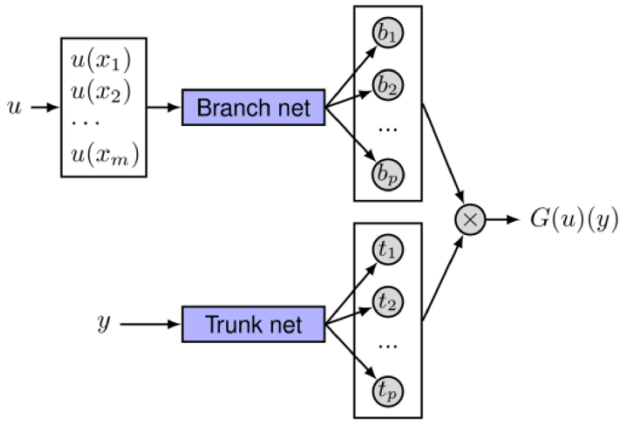
$$G: U \longrightarrow S$$
$$u(x) \to s(y)$$

$$\begin{array}{ccc} L^2(\mathcal{X}) & \xrightarrow{\mathcal{G}} & L^2(\mathcal{Y}) \\ \downarrow \mathcal{E} & & \uparrow \mathcal{D} \\ \mathbb{R}^m & \xrightarrow{\mathcal{A}} & \mathbb{R}^n \end{array}$$

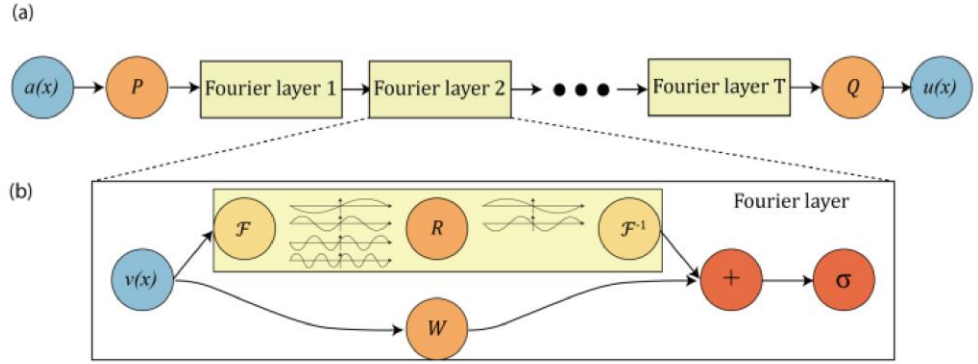
Application: Learning the solution operator of parametric PDEs



DeepONet

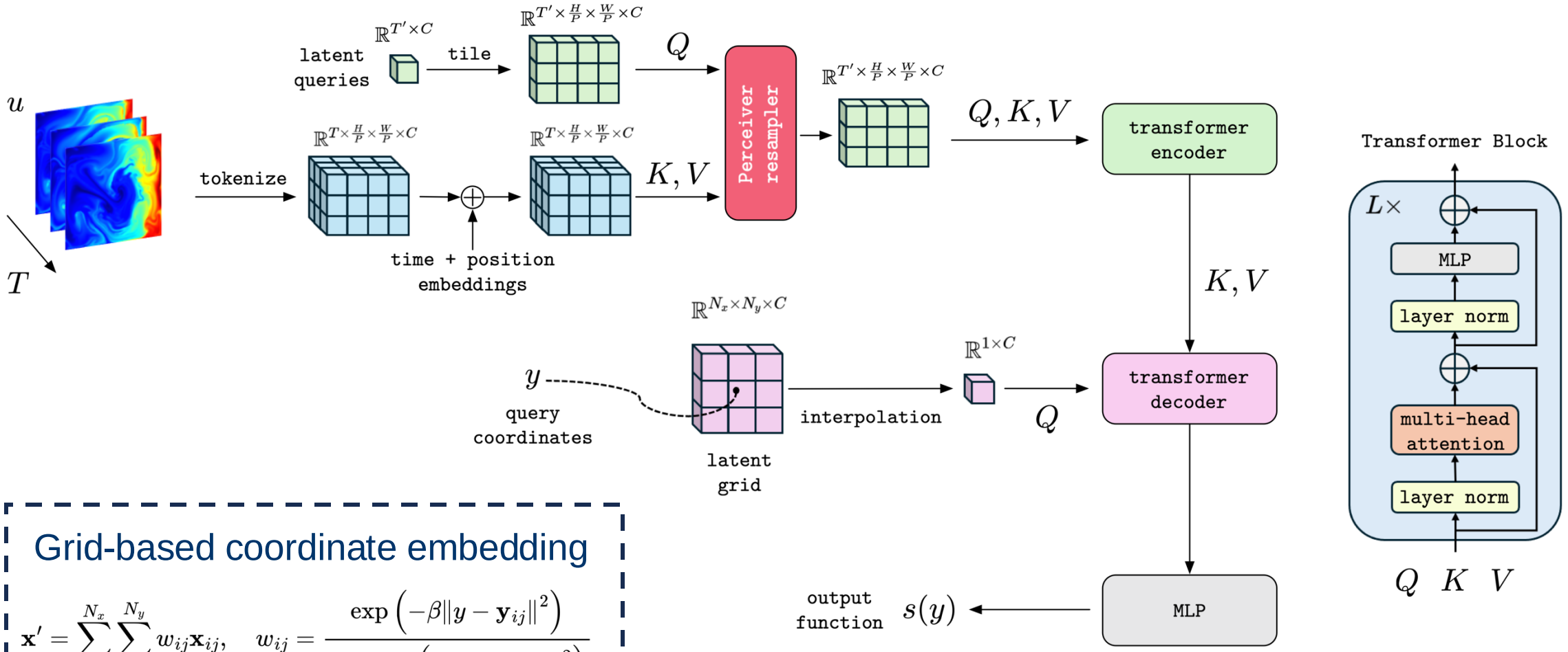


Fourier Neural Operator (FNO)



1. Lu, Lu, et al. "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators." Nature machine intelligence 3.3 (2021): 218-229.
2. Li, Zongyi, et al. "Fourier neural operator for parametric partial differential equations." arXiv preprint arXiv:2010.08895 (2020).
3. Seidman, Jacob, et al. "NOMAD: Nonlinear manifold decoders for operator learning." Advances in Neural Information Processing Systems 35 (2022): 5601-5613.

Continuous Vision Transformer (CViT)



Grid-based coordinate embedding

$$\mathbf{x}' = \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} w_{ij}\mathbf{x}_{ij}, \quad w_{ij} = \frac{\exp\left(-\beta\|y-\mathbf{y}_{ij}\|^2\right)}{\sum_{ij}\exp\left(-\beta\|y-\mathbf{y}_{ij}\|^2\right)}$$

Lipschitz constant & Spectral Bias

Lipschitz constant: $Lip(f) = \sup_{x \neq y} \frac{|f(y) - f(x)|}{|y - x|}$

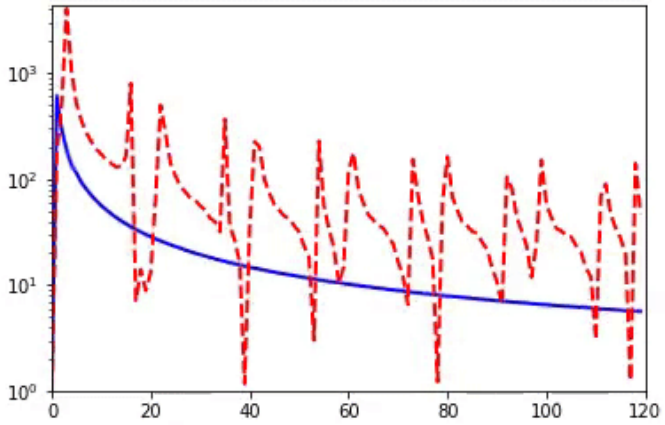
MLP: $\mathbf{f}_\theta^{(l)}(x) = \mathbf{W}^{(l)} \cdot \mathbf{g}^{(l-1)}(x) + \mathbf{b}^{(l)}, \quad \mathbf{g}^{(l)}(x) = \sigma\left(\mathbf{f}_\theta^{(l)}(x)\right), \quad l = 1, 2, \ldots, L,$

$$f_\theta(x) = \mathbf{W}^{(L+1)} \cdot \mathbf{g}^{(L)}(x) + \mathbf{b}^{(L+1)}, \quad \mathbf{W}^{(l)} \sim \mathcal{N}\left(0, \frac{2}{d_l + d_{l-1}}\right)$$

Assume $\dot{\sigma}(\cdot) \in [0, 1]$

$$\begin{aligned} \operatorname{Lip}(f_\theta) &\leq \sup_{\mathbf{z}^{(i-1)} \in [0,1]^n} \left\| \prod_{i=L+1}^{2} \left(\mathbf{W}^{(i)} \cdot \operatorname{diag}\left(\mathbf{z}^{(i-1)}\right)\right) \cdot \mathbf{W}^{(1)} \right\|_2 \\ &\leq \prod_{i=1}^{L+1} \left\| \mathbf{W}^{(i)} \right\|_2. \end{aligned}$$

Spectral bias



1.Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y. and Courville, A., 2019, May. On the spectral bias of neural networks. In International conference on machine learning (pp. 5301-5310). PMLR.

2. Latorre, F., Rolland, P. and Cevher, V., 2020. Lipschitz constant estimation of neural networks via sparse polynomial optimization. arXiv preprint arXiv:2004.08688.

Lipschitz constant of Coordinate Embeddings

Trivial coordinate embedding $\phi_1(x) = \mathbf{W}x + b, \quad \mathbf{W} \sim \mathcal{N}\left(0, \frac{1}{d}\right)$

$$\mathbb{E}\,\mathrm{Lip}\,(\phi_1) = \frac{\sqrt{2}}{\sqrt{d}} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \approx 1 - \frac{1}{2d}$$

Random Fourier features $\phi_2(x) = \begin{bmatrix} \cos(\mathbf{W}x) \\ \sin(\mathbf{W}x) \end{bmatrix}, \quad \mathbf{W} \sim \mathcal{N}(0, s)$

$$\mathbb{E}\,\mathrm{Lip}\,(\phi_2) = 2s \cdot \frac{\Gamma\left(\frac{d/2+1}{2}\right)}{\Gamma(d/4)} \approx s\sqrt{d}\left(1 - \frac{1}{8d}\right)$$

Grid-based embedding $\phi_\beta(x) = \sum_{i=1}^{n} \omega_i(x)\mathbf{f}_i = \sum_{i=1}^{n} e^{-\beta(x-x_i)^2}\mathbf{f}_i, \quad \|\mathbf{f}_i\| \leq M$

$$M\sqrt{2\beta}e^{-\frac{1}{2}} \lesssim \mathrm{Lip}\,(\phi_\beta) \lesssim 2M\sqrt{2\beta}e^{-\frac{1}{2}}$$

Tancik, Matthew, et al. "Fourier features let networks learn high frequency functions in low dimensional domains." Advances in neural information processing systems 33 (2020): 7537-7547.

Advection of Discontinuous Waves

Problem setup

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \quad x \in [0,1)$$
$$u(0) = u_0$$

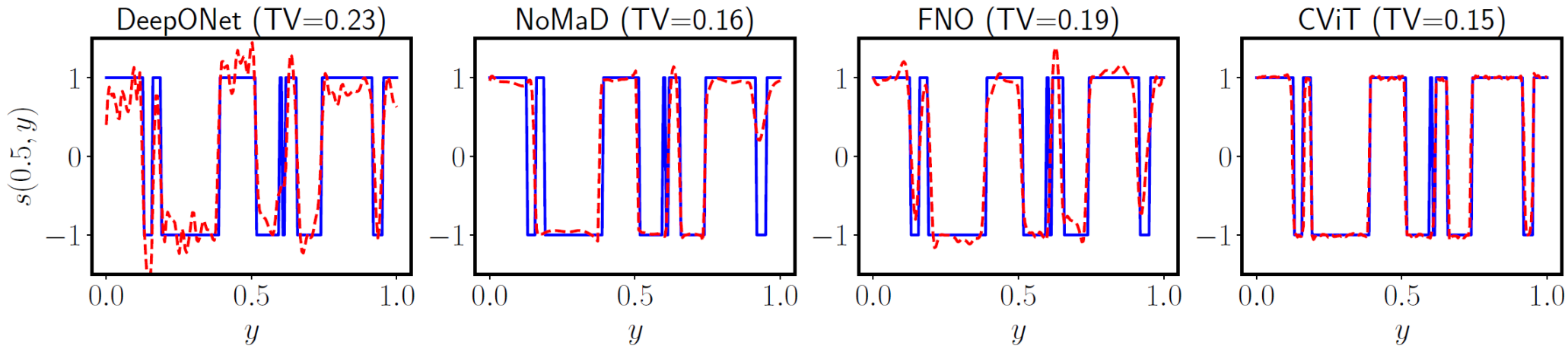
Initial condition

$$u_0 = -1 + \mathbf{21}_{\{\tilde{u_0} \geq 0\}}$$

$$\widetilde{u_0} \sim \mathcal{N}(0, C) \text{ and, } \mathrm{C} = \left(-\Delta + \tau^2\right)^{-d}$$

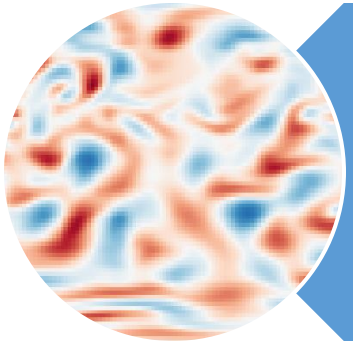
Goal: Learning the mapping from initial condition to $u(0.5,\cdot)$

		Rel. L2 error (%, ↓)			Total Variation error (↓)		
Model	**# Params**	**Mean**	**Median**	**Worst-case**	**Mean**	**Median**	**Worst-case**
FNO	4.98M	13.94	12.03	56.52	0.0406	0.0309	0.1910
DeepONet	2.20M	13.77	12.20	39.64	0.0572	0.0453	0.2300
NoMaD	2.21M	13.64	11.01	69.00	**0.0326**	**0.0245**	0.1732
CViT	3.03M	**11.70**	**10.68**	**35.44**	0.0334	0.0274	**0.1516**



de Hoop, M.V., Huang, D.Z., Qian, E. and Stuart, A.M., 2022. The cost-accuracy trade-off in operator learning with neural networks. arXiv preprint arXiv:2203.13181.

Benchmarks



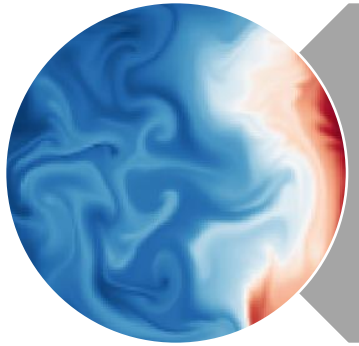
Shallow Water

Trajectories: 5,600

Time steps: 11

Resolution: 96 x 192

Variables: Vorticity & Pressure



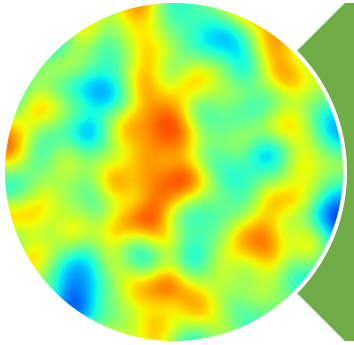
Incompressible Navier-Stokes

Trajectories: 6,500

Time steps: 14

Resolution: 128 x 128

Variables: Passive Scalar & Velocity



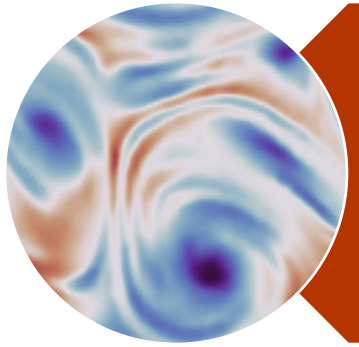
Diffusion Reaction

Trajectories: 900

Time steps: 100

Resolution: 128 x 128

Variables: Density fields



Compressible Navier-Stokes

Trajectories: 9,000

Time steps: 100

Resolution: 128 x 128

Variables: Passive Scalar & Velocity

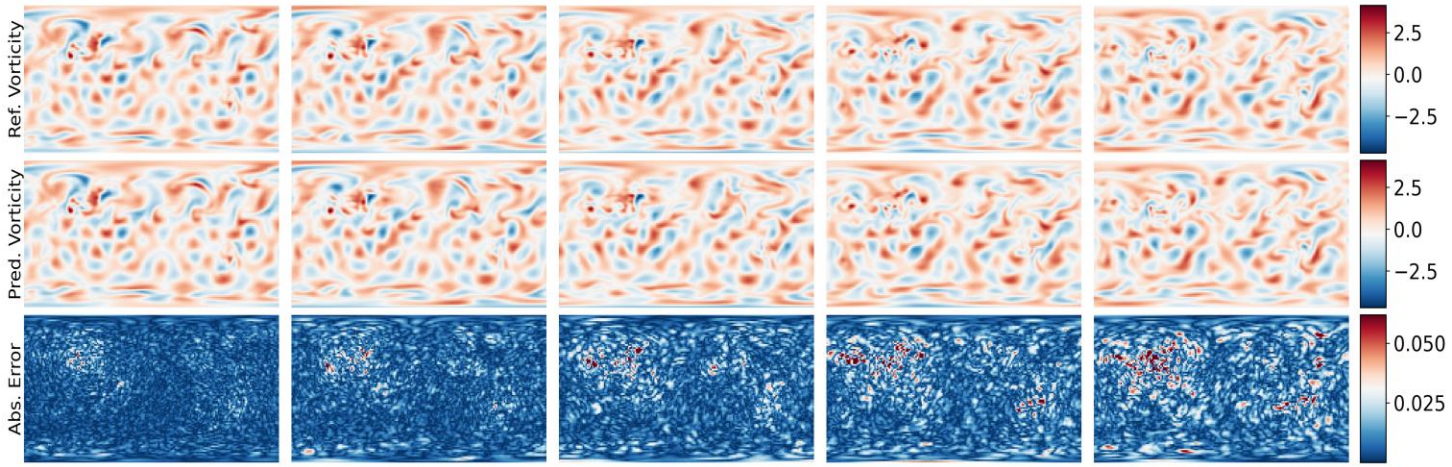
1. Takamoto, M., Praditia, T., Leiteritz, R., MacKinlay, D., Alesiani, F., Pflüger, D. and Niepert, M., 2022. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, *35*, pp.1596-1611.

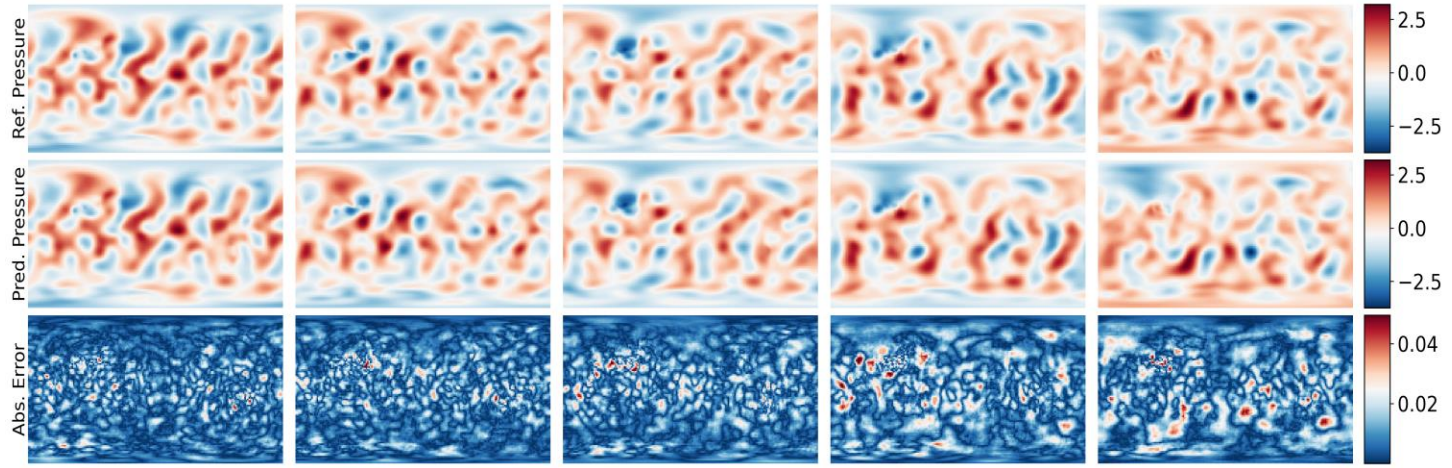
2.Gupta, J.K. and Brandstetter, J., 2022. Towards multi-spatiotemporal-scale generalized pde modeling. arXiv preprint arXiv:2209.15616.

Shallow Water Equation

Model	# Params	Rel. L^2 error (↓)
DilResNet	4.2 M	13.20%
U-Net$_{att}$	148 M	5.684%
FNO	268 M	3.97 %
U-F2Net	344 M	1.89 %
UNO	440 M	3.79%
CViT-S	13 M	**4.47 %**
CViT-B	30 M	**2.69 %**
CViT-L	92 M	**1.56 %**

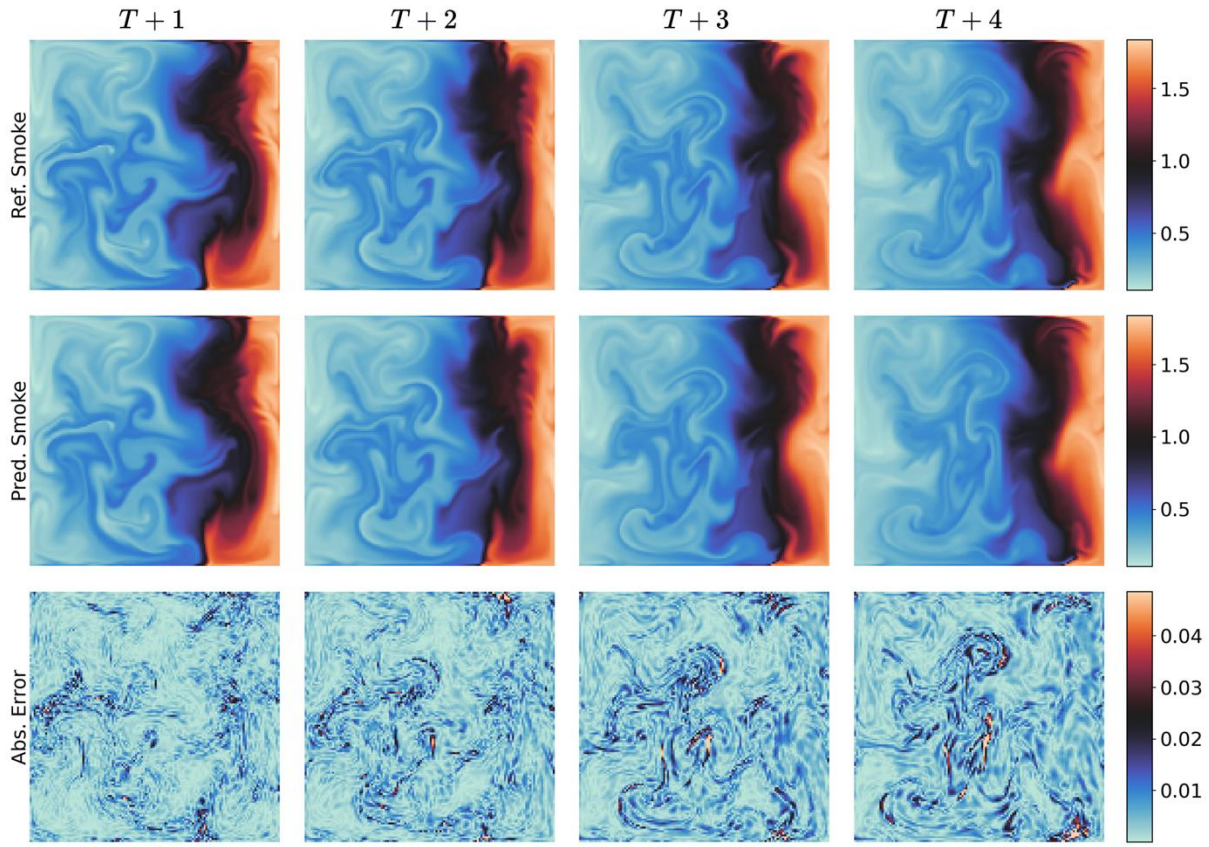
Model	CViT-S	CViT-B	CViT-L
Encoder layers	5	10	15
Embedding dim	384	512	768
MLP width	384	512	1536
Heads	6	8	12
# Params	13 M	30 M	92 M





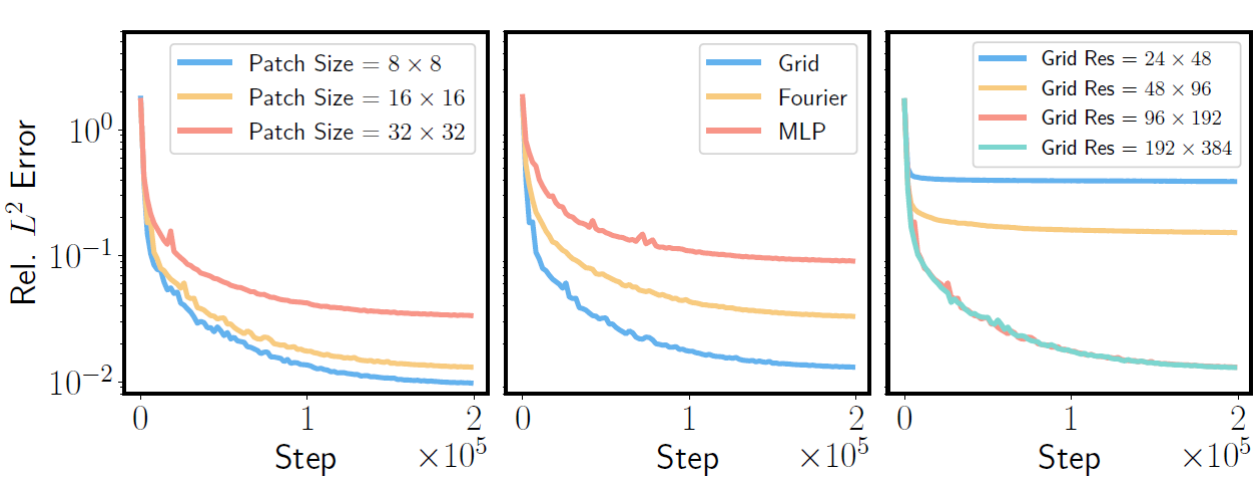
Results

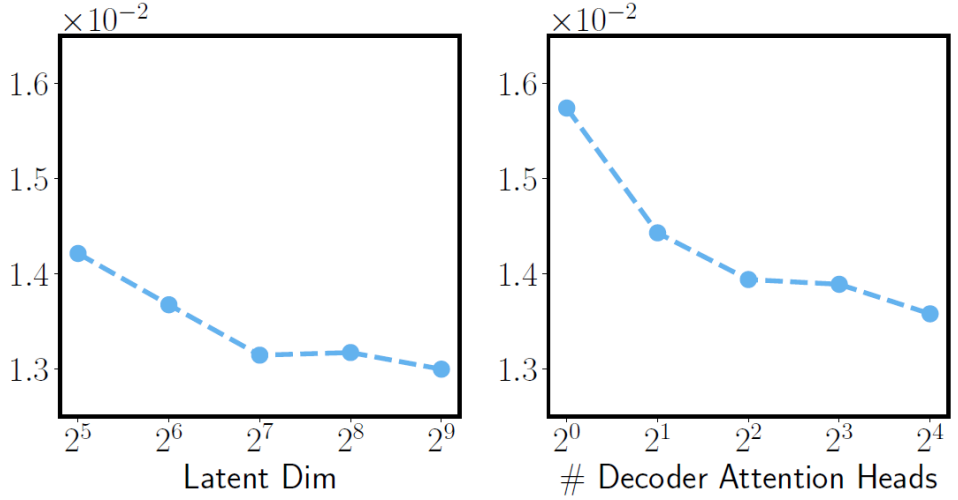
Model	# Params	NS	CNS	DR
FNO	0.5 M	9.12 %	9.60 %	12.00 %
FFNO	1.3 M	8.39 %	5.20 %	5.71 %
GK-T	1.6 M	9.52%	3.77%	3.59 %
GNOT	1.8 M	17.20 %	4.20 %	3.11 %
Oformer	1.9 M	13.50 %	6.25 %	1.92 %
DPOT-Ti	7 M	12.50 %	3.97 %	3.21 %
MPP-S	30M	-	-	1.12 %
DPOT-S	30 M	9.91 %	3.37 %	3.79 %
MPP-L (Pre-trained)	400 M	-	-	0.98 %
DPOT-L (Pre-trained)	500 M	7.98 %	2.16 %	2.32 %
DPOT-L (Fine-tuned)	500 M	2.78 %	1.31 %	0.73 %
DPOT-H (Pre-trained)	1.03 B	3.79 %	1.80 %	1.91 %
CViT-S	13 M	**3.75 %**	**2.71 %**	**1.13 %**
CViT-B	30 M	**3.18 %**	**1.99 %**	**1.11 %**
CViT-L	92 M	**2.35 %**	**1.29 %**	**0.68 %**



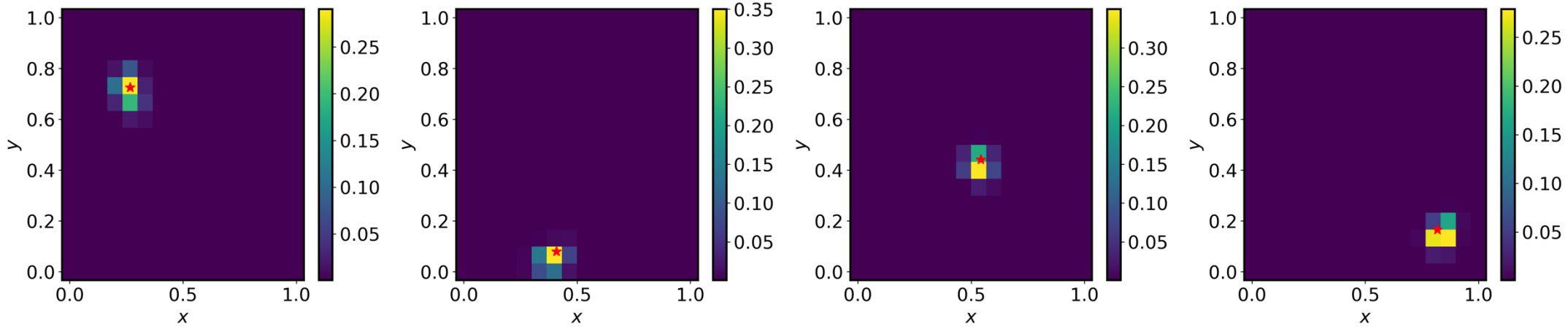
Hao, Z., Su, C., Liu, S., Berner, J., Ying, C., Su, H., Anandkumar, A., Song, J. and Zhu, J., 2024. Dpot: Auto-regressive denoising operator transformer for large-scale pde pre-training. arXiv preprint arXiv:2403.03542.

Ablation Studies





Visualizations of decoder cross-attention map



Thank you!

Penn
University of Pennsylvania

ICLR
International Conference On
Learning Representations

Yale