SAOT: An Enhanced Locality-Aware Spectral Transformer for Solving PDEs

Chenhong Zhou[1], **Jie Chen**[1†], **Zaifeng Yang**[2]

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
20482795@life.hkbu.edu.hk, chenjie@comp.hkbu.edu.hk, yang_zaifeng@a-star.edu.sg

Abstract

Neural operators have shown great potential in solving a family of Partial Differential Equations (PDEs) by modeling the mappings between input and output functions. Fourier Neural Operator (FNO) implements global convolutions via parameterizing the integral operators in Fourier space. However, it often results in over-smoothing solutions and fails to capture local details and high-frequency components. To address these limitations, we investigate incorporating the spatial-frequency localization property of Wavelet transforms into the Transformer architecture. We propose a novel Wavelet Attention (WA) module with linear computational complexity to efficiently learn locality-aware features. Building upon WA, we further develop the Spectral Attention Operator Transformer (SAOT), a hybrid spectral Transformer framework that integrates WA's localized focus with the global receptive field of Fourier-based Attention (FA) through a gated fusion block. Experimental results demonstrate that WA significantly mitigates the limitations of FA and outperforms existing Wavelet-based neural operators by a large margin. By integrating the locality-aware and global spectral representations, SAOT achieves state-of-the-art performance on six operator learning benchmarks and exhibits strong discretization-invariant ability.

Code — https://github.com/chenhong-zhou/SAOT

Introduction

Solving partial differential equations (PDEs) is a fundamental and important task in real-world applications, enabling people to understand complex physical systems and make reliable decisions (Roubíček 2013; Azizzadenesheli et al. 2024). Recently, deep learning has emerged as a promising tool for learning solution operators for PDEs, mainly attributed to its powerful modeling ability. Versatile neural operators have been trained to learn the mappings between input and output function spaces, and then they can generalize well to inputs unseen during training (Li et al. 2020; Lu et al. 2021). This learning paradigm facilitates much faster inference than traditional numerical methods and shows great potential in surrogate modeling of PDEs (Li et al. 2021; Hao et al. 2024; Yue, Zhu, and Yang 2024). Specifically, as a representative architecture of neural operators, Fourier neural operator (FNO) (Li et al. 2021) parameterizes the integral kernel in Fourier space. This pioneering work has achieved impressive results in a wide range of applications (Jiang et al. 2023; Liu and Tang 2024; Li et al. 2024; Cho et al. 2024; Leng et al. 2025; Liu et al. 2025) and sparked the surge of a series of follow-up methods like Geo-FNO (Li et al. 2023), U-FNO (Wen et al. 2022), F-FNO (Tran et al. 2023), AFNO (Guibas et al. 2022), etc. Typically, FNO and its variants compute a forward Fast Fourier Transform (FFT), perform a matrix multiplication between Fourier modes and a learnable weight tensor, and conduct an inverse FFT, which amounts to global convolution in the spatial domain according to the convolution theorem (Li et al. 2021, 2023; Poli et al. 2022). Hence, these methods can capture long-range dependencies by exploiting the strengths of learning in the frequency domain. However, such global convolution operations often lead to over-smoothing solutions and have deficiencies in capturing local and high-frequency details (Bachman, Narici, and Beckenstein 2000; Tripura and Chakraborty 2022; Oommen et al. 2025), which exacerbates the spectral bias phenomenon (Rahaman et al. 2019; Cao et al. 2021). Some works have tried to replace Fourier transform with Wavelet transform to better capture local variations and nuances, such as MWT (Gupta, Xiao, and Bogdan 2021) and WNO (Tripura and Chakraborty 2022). Wavelet basis functions, being localized in both space and frequency, are well-suited for capturing intricate local details and handling spatially discontinuous or abrupt signals, which are capabilities that the Fourier transform inherently lacks (Bachman, Narici, and Beckenstein 2000; Khan and Yener 2018). Nevertheless, current implementations of Wavelet-based neural operators fail to achieve expected approximation accuracy, a shortcoming becoming significantly apparent against the rising dominance of Transformer-based models.

As a powerful architecture, Transformer (Vaswani et al. 2017) has attracted growing interest and has been progressively adopted in operator learning (Guibas et al. 2022; Kovachki et al. 2023; Hao et al. 2024). The core of Transformers is the self-attention mechanism for effective token mixing to learn the relationships among tokens (Guibas et al. 2022; Lee-Thorp et al. 2021). However, the standard dot-

†Corresponding author.

product self-attention makes it computationally infeasible to model intricate correlations of massive mesh points, due to its quadratic complexity with respect to the sequence length (Hao et al. 2023; Wu et al. 2024). Recently, Fourier transforms have been integrated into Transformer architectures to reduce computational complexity, such as FNet (Lee-Thorp et al. 2021), GFNet (Rao et al. 2021), and AFNO (Guibas et al. 2022). These different Fourier-based attentions have been proposed to replace standard self-attention and enable token mixing in the Fourier domain, but the aforementioned limitations still exist. To mitigate these issues, we investigate introducing the Wavelet transform into a Transformer architecture and propose a novel Wavelet-based attention mechanism. To further combine the complementary benefits of the Fourier and Wavelet transforms, we propose a hybrid spectral Transformer framework that enhances the spectral representation for better approximation accuracy.

In this work, we propose a Wavelet Attention with linear complexity, aiming to capture local and high-frequency details while reducing computational expenses. Wavelet attention employs Fast Wavelet Transform (FWT) to decompose the input data into a sequence of wavelet subbands (coefficients), then applies a linearized attention mechanism between these wavelet subbands, and finally does inverse FWT. To fully harness the respective advantages of both Wavelet and Fourier transforms, we introduce a Spectral Attention Operator Transformer (SAOT) with the hybrid spectral attention. Specifically, a spectral attention layer consists of a Wavelet attention (WA) and a parallel Fourier attention (FA) as well as a gated fusion block to adaptively merge locality-aware representations from WA and global features from FA, thereby greatly boosting performance. Overall, our contributions are summarized as follows:

- We propose WA with linear complexity to learn locality-aware representations by triggering self-attention learning between wavelet frequency subbands, which can compensate for the spectral deficiencies of FA.
- We propose SAOT with Spectral Attention to adaptively adjust the contributions of global features from FA and local details from WA, which leverages the complementary advantages of two different frequency-domain operations.
- Experimental results demonstrate that WA mitigates the loss of high-frequency details and outperforms current Wavelet-based neural operators by a large margin. By combining WA and FA, SAOT achieves state-of-the-art performance on several standard benchmarks and also presents an excellent discretization-invariant property.

Related Work

Neural operators aim to learn mappings between infinite-dimensional function spaces and therefore they are mesh-invariant, i.e., agnostic to the discretization of input and output functions (Lu et al. 2021; Li et al. 2021; Kovachki et al. 2023). With the exception of some earlier works, such as DeepONet (Lu, Jin, and Karniadakis 2019) and GNO (Li et al. 2020), current prevailing neural operator approaches can be broadly categorized into the following two classes.

Spectral-based Neural Operators As a seminal work, FNO (Li et al. 2021) pioneered learning operator mappings in spectral space, inspiring numerous works based on typical spectral techniques. Applying the convolution theorem, FNO parameterizes the kernel integral operator by computing a matrix multiplication between Fourier modes and a learnable weight, which is equivalent to performing global convolution in the spatial domain (Li et al. 2021; Guibas et al. 2022). Subsequently, U-NO (Rahman, Ross, and Azizzadenesheli 2023) and U-FNO (Wen et al. 2022) plug FNO into U-Net (Ronneberger, Fischer, and Brox 2015) to improve the performance. Geo-FNO (Li et al. 2023) extends the applicability of FNO from uniform grids to arbitrary geometries. F-FNO (Tran et al. 2023) factorizes the Fourier representation to process each spatial dimension independently. SNO (Fanaskov and Oseledets 2022) aims to reduce aliasing errors in FNO using basic spectral methods. However, global convolution implemented in FNO and its variants is often vulnerable to over-smoothing and neglects local features (Tripura and Chakraborty 2022; Liu-Schiaffini et al. 2024). To implement local receptive fields, local neural operators (Liu-Schiaffini et al. 2024) are developed to learn integral operators with locally supported kernels. In addition, some works have utilized Wavelet transforms to better capture locality by introducing multiscale wavelet bases in MWT (Gupta, Xiao, and Bogdan 2021) and wavelet integral layers in WNO (Tripura and Chakraborty 2022).

Attention-based Neural Operators Another class of operator learning approaches is based on Transformer architectures, as the attention mechanism provides large modeling capability and flexibility (Hao et al. 2023). The quadratic complexity of attention is a critical challenge when handling large-scale irregular grids in PDEs. Much effort has been devoted to tackling the complexity issue by adopting smaller patches during patchify operation (Wu et al. 2023; Wang et al. 2024; Hagnberger et al. 2024), introducing Fourier-based attentions (Hao et al. 2024; Guibas et al. 2022; Rao et al. 2021; Lee-Thorp et al. 2021), and designing efficient attentions (Cao 2021; Li, Meidani, and Farimani 2023; Li, Shu, and Farimani 2023; Hao et al. 2023; Chen and Wu 2024). Specifically, AFNO (Guibas et al. 2022) adapts FNO as an efficient token mixer with quasi-linear complexity, which is applied for operator learning in DPOT (Hao et al. 2024). Galerkin-type attention (Cao 2021) is developed to remove softmax normalization and achieve linear scaling. Afterwards, OFormer (Li, Meidani, and Farimani 2023), GNOT (Hao et al. 2023), and ONO (Xiao et al. 2024) employ the linear-complexity variants of attention (Kitaev, Kaiser, and Levskaya 2020; Katharopoulos et al. 2020a; Choromanski et al. 2021) for learning operators. Furthermore, IPOT (Lee and Oh 2024), PiT (Chen and Wu 2024), and Transolver (Wu et al. 2024) design novel attention mechanisms to improve model performance with affordable computational complexity.

Method

To compensate for the deficiencies in Fourier-based attention, we first propose an efficient Wavelet attention that

learns locality-aware features to enhance high-frequency details while reducing computational complexity. Next, we introduce a spectral attention layer to merge the features from a WA and a parallel FA via a gated fusion block. Finally, we present the overall architecture of SAOT and analyze its computational complexity.

Problem Setup Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. The input and output functions a and u are drawn from Banach spaces $\mathcal{A}$ and $\mathcal{U}$, respectively: $a \in \mathcal{A}(\Omega; \mathbb{R}^{d_a}), u \in \mathcal{U}(\Omega; \mathbb{R}^{d_u})$. Suppose Φ is an operator mapping from input to output function, i.e., $\Phi : \mathcal{A} \to \mathcal{U},\ a \mapsto u$. In practice, representing the function directly is hard. Hence, training data are the values of the input and output functions sampled on regular or irregular meshes g. Our goal is to learn a neural operator Φ_θ to approximate Φ from a set of observed input-output pairs, where θ denotes the trainable parameters.

Preliminary: Self-attention Given the input sequence $X = \{\boldsymbol{x}_j\}_{j=1}^N \in \mathbb{R}^{N \times D}$ which contains N elements and each element has D feature channels, self-attention employs three trainable matrices $W^q, W^k, W^v \in \mathbb{R}^{D \times D}$ to obtain the corresponding representations (Q: queries, K: keys, and V: values):

$$\boldsymbol{q}_j = \boldsymbol{x}_j W^q, \quad \boldsymbol{k}_j = \boldsymbol{x}_j W^k, \quad \boldsymbol{v}_j = \boldsymbol{x}_j W^v, \quad (1)$$

where $\boldsymbol{q}_j$, $\boldsymbol{k}_j$, and $\boldsymbol{v}_j$ are the j-th rows of Q, K, and V, respectively. Softmax attention is a specific form of self-attention, where the similarity score is computed as the exponential of a dot-product between queries and keys, which can be formulated as follows:

$$\boldsymbol{s}_i = \frac{\sum_{j=1}^N \exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_j / \tau) \boldsymbol{v}_j}{\sum_{l=1}^N \exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_l / \tau)}, \quad (2)$$

where τ is a temperature hyperparameter and $\boldsymbol{s}_i$ denotes the i-th row vector of attention matrix $S \in \mathbb{R}^{N \times D}$. The computational cost of self-attention is scaling quadratically w.r.t. the sequence length, i.e., $\mathcal{O}\left(N^2\right)$. A generalized self-attention function for any similarity function (Katharopoulos et al. 2020b) can be expressed as:

$$\boldsymbol{s}_i' = \frac{\sum_{j=1}^N \mathrm{sim}\left(\boldsymbol{q}_i, \boldsymbol{k}_j\right) \boldsymbol{v}_j}{\sum_{l=1}^N \mathrm{sim}\left(\boldsymbol{q}_i, \boldsymbol{k}_l\right)}, \quad (3)$$

If the similarity function is $\mathrm{sim}\left(\boldsymbol{q}, \boldsymbol{k}\right) = \exp\left(\boldsymbol{q} \cdot \boldsymbol{k} / \tau\right)$, the generalized form in Eq. 3 reduces to the softmax attention in Eq. 2. Previous works have demonstrated that a kernel with non-negative similarity scores can be used to replace the exponential kernel in softmax attention and linearize the attention to reduce the computational complexity (Cao 2021; Choromanski et al. 2021; Katharopoulos et al. 2020b).

Wavelet Attention (WA)

An important benefit of the Wavelet transform is providing spatial-frequency localized representations; thus, we propose the Wavelet Attention, an efficient implementation for mixing tokens in the Wavelet domain. Formally, given a mesh set with N points , we consider $\Omega \subset \mathbb{R}^2$ as a depicted example and thus $N = H \times W$. The inputs are projected into deep feature maps $X \in \mathbb{R}^{H \times W \times D}$. We first embed X into $\bar{X} \in \mathbb{R}^{H \times W \times \frac{D}{4}}$ via a convolution layer to reduce channel dimension before employing Fast Wavelet Transform (FWT). Then we use FWT to decompose $\bar{X}$ into different wavelet subbands (coefficients) with common smaller sizes and different frequency components. Here Haar wavelet is chosen for FWT as in (Liu et al. 2020; Yao et al. 2022) for simplicity. Concretely, the high-pass filter $f_H = (1/\sqrt{2}, -1/\sqrt{2})$ and low-pass filter $f_L = (1/\sqrt{2}, 1/\sqrt{2})$ in FWT are applied along rows to transform $\bar{X}$ into subbands X_H and X_L. Then, these filters are applied to the columns of X_H and X_L, resulting in four distinct wavelet subbands: X_{HH}, X_{HL}, X_{LH}, and X_{LL}, where each subband has dimensions of $\mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times \frac{D}{4}}$. X_{LL} denotes the low-frequency component in both dimensions, capturing fundamental features at a coarse level. X_{LH}, X_{HL}, and X_{HH} signify the high-frequency components, preserving the details at a fine level. These wavelet subbands collectively form a compact and complete representation of the original data at different levels of detail and orientation. Hence, we concatenate them along the channel dimension to form $\widetilde{X} = \mathrm{Concat}(X_{LL}, X_{LH}, X_{HL}, X_{HH}) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times D}$. Notice that the spatial dimensions are halved in both height and width, resulting in a one-quarter reduction in total, which lowers computational demands for the subsequent attention calculation.

By leveraging a kernel feature map $\phi\left(x\right)$ to linearize the attention mechanism, we can rewrite Eq. 3 and further simplify it via the associative property of matrix product:

$$\boldsymbol{s}_i' = \frac{\sum_{j=1}^N \phi\left(\boldsymbol{q}_i\right)^T \phi\left(\boldsymbol{k}_j\right) \boldsymbol{v}_j}{\sum_{l=1}^N \phi\left(\boldsymbol{q}_i\right)^T \phi\left(\boldsymbol{k}_l\right)} = \frac{\phi\left(\boldsymbol{q}_i\right)^T \left(\sum_{j=1}^N \phi\left(\boldsymbol{k}_j\right) \otimes \boldsymbol{v}_j\right)}{\phi\left(\boldsymbol{q}_i\right)^T \sum_{l=l}^N \phi\left(\boldsymbol{k}_l\right)}. \quad (4)$$

According to Eq. 4, $\sum_{j=1}^N \phi\left(\boldsymbol{k}_j\right) \otimes \boldsymbol{v}_j$ and $\sum_{l=1}^N \phi\left(\boldsymbol{k}_l\right)$ can be computed once and efficiently reused across all queries, so that such an attention has efficient computational and memory complexity $\mathcal{O}\left(N\right)$. Here we use the feature map as in (Katharopoulos et al. 2020b): $\phi\left(x\right) = \mathrm{elu}(x) + 1$, where $\mathrm{elu}(\cdot)$ denotes the exponential linear unit activation function (Clevert 2015). Finally, we use $\mathrm{LinearAttn}$ to represent this linearized attention. Before performing self-attention learning over wavelet subbands, we optionally apply a convolution operation like a kernel of 3×3 to $\widetilde{X}$ to enforce spatial locality, thereby generating the locally contextualized features X^w (Yao et al. 2022). Next, X^w is reshaped and linearly transformed into queries/keys/values: $Q^w, K^w, V^w \in \mathbb{R}^{n \times D}$, where $n = \frac{H}{2} \times \frac{W}{2}$. Then we employ the linearized attention mechanism to capture the long-range contextualized information among wavelet subbands: $S' = \mathrm{LinearAttn}(Q^w, K^w, V^w)$.

Afterwards, we reshape S' into $\mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times D}$ and apply the inverse FWT (IFWT) to obtain the reconstruction $X^r \in \mathbb{R}^{H \times W \times \frac{D}{4}}$. Finally, the reconstructed and original data are concatenated, followed by a linear embedding layer to compose the final output of WA: i.e. $X^{\mathrm{WA}} = \mathrm{Linear}\left(\mathrm{Concat}(X, X^r)\right)$. As a result, the main operations in the WA can be summarized

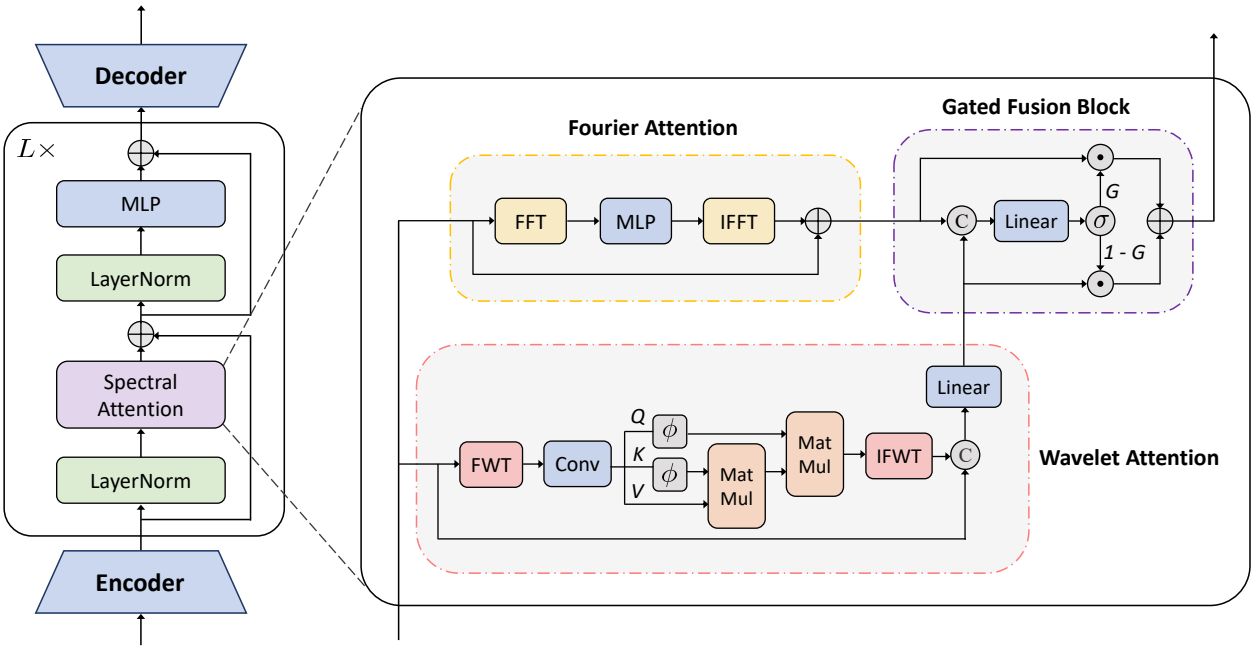

Figure 1: Architecture of the SAOT model. It contains three main components: an encoder, a processor with L pre-norm Transformer blocks, and a decoder. The core of our Transformer block is the proposed spectral attention, which is composed of a Fourier Attention (FA) and a Wavelet Attention (WA), as well as a gated fusion block.

as FWT-Conv-LinearAttn-IFWT. This attention mechanism harnesses the benefits of wavelets to accomplish local receptive fields and significantly reduce the complexity.

Spectral Attention Layer

We propose a spectral attention layer that integrates a WA and a FA in parallel, combined with a gated fusion block to synergize the advantages of Wavelet and Fourier transforms.

Fourier Attention (FA) Here, we utilize the FA similar to AFNO (Guibas et al. 2022), which adapts FNO for token mixing in Transformer architectures. Unlike AFNO, our FA omits sparsity operations to retain all frequency modes, thereby preserving the full expressivity of the representation (Hao et al. 2024). Concretely, FA performs the global convolution by approximating the kernel integral operator in Fourier space (Li et al. 2021; Guibas et al. 2022):

$$\mathcal{K}(X)(g) = \mathcal{F}^{-1}(R_\psi \cdot \mathcal{F}(X))(g), \qquad \forall g \in \Omega \quad (5)$$

where $\mathcal{K}$ is a kernel integral operator parameterized by R_ψ, $\mathcal{F}$, $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse, and X denotes the input function discretized on mesh g. Specifically, R_ψ represents block-wise MLP layers instead of a learnable complex-valued weight tensor in FNO to dramatically reduce training parameters and memory (Guibas et al. 2022). Thus, the FA can be summarized as FFT-MLP-IFFT, and its output is denoted as X'. Following AFNO, we add a residual term to X' to generate the final output of FA: $X^{\text{FA}} = X + X'$. The FA module is running in parallel to the proposed WA.

Gated Fusion Block To compensate for local details, we incorporate a WA to complement FA and introduce a gated fusion block to adaptively merge the features from two spectral operations. Concretely, given the features $X^{\text{FA}} \in \mathbb{R}^{N\times D}$ and $X^{\text{WA}} \in \mathbb{R}^{N\times D}$ derived from FA and WA respectively, we first concatenate them along the channel dimension and then feed them into a linear embedding layer, followed by a sigmoid activation function:

$$G = \sigma\left(\text{Linear}\left(\text{Concat}(X^{\text{FA}}, X^{\text{WA}})\right)\right), \quad (6)$$

where $\sigma(\cdot)$ is the sigmoid activation function, and $G \in \mathbb{R}^{N\times D}$ denotes the generated gating map. Next, this gating map can be used to fuse the features by performing an element-wise weighted sum:

$$X^{\text{SA}} = G \odot X^{\text{FA}} + (1 - G) \odot X^{\text{WA}}, \quad (7)$$

where $\odot$ denotes element-wise multiplication, and X^{SA} denotes the fused features, i.e., the final output of a spectral attention (SA) layer. As a result, the SA layer integrates global dependencies from a FA and local features from a WA to leverage the complementary advantages of these two frequency-domain operations.

Spectral Attention Operator Transformer (SAOT)

In general, our SAOT model is built upon a Transformer architecture where the standard softmax attention is replaced with the proposed spectral attention layer.

Overall Architecture Specifically, SAOT adopts the typical Encoder-Processor-Decoder architecture:

$$\Phi_\theta = \text{Decoder} \circ \text{LAYER}_L \circ \cdots \circ \text{LAYER}_1 \circ \text{Encoder}, \quad (8)$$

Benchmarks	Input	Output	Geometry	N	N_t	Train set	Test set
Darcy	Diffusion coefficient	Fluid pressure	Regular grid	85×85	-	1000	200
NS	Past velocity	Future velocity	Regular grid	64×64	10	1000	200
Airfoil	Mesh points	Mach number	Structured mesh	221×51	-	1000	200
Pipe	Mesh points	Fluid velocity	Structured mesh	129×129	-	1000	200
Plasticity	Mesh points	Mesh deformation	Structured mesh	101×31	20	900	80
Elasticity	Structure	Inner stress	Point cloud	972	-	1000	200

Table 1: Summary of benchmarks. N means the spatial resolution, and N_t means temporal dimension. The last two columns denote the number of samples in the train and test sets.

where the encoder is usually implemented by linear layers to lift the input function from $\mathbb{R}^{d_a}$ to a high-dimensional feature space $\mathbb{R}^D$. As shown in Figure 1, the processor is composed of L stacked pre-norm Transformer blocks (Vaswani et al. 2017; Xiong et al. 2020), and the l-th block can be formalized as follows:

$$\begin{aligned} \hat{X}^l &= \mathrm{SA}\left(\mathrm{LN}\left(X^{l-1}\right)\right) + X^{l-1} \\ X^l &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{X}^l\right)\right) + \hat{X}^l, \end{aligned} \tag{9}$$

where $l = 1, \cdots, L$, LN denotes layer normalization (Ba, Kiros, and Hinton 2016), and MLP is usually implemented with a two-layer feedforward network. $X^l \in \mathbb{R}^{N \times D}$ is the output of the l-th block. $X^0 \in \mathbb{R}^{N \times D}$ denotes the input deep features from the encoder. Finally, X^L is further fed to the decoder to perform a linear projection to the output dimension $\mathbb{R}^{d_u}$.

Computational Complexity Since a SA is composed of a FA and a WA running in parallel, its overall complexity is dominated by the higher-complexity component. Assuming $N \gg D$, the FA has the complexity of $\mathcal{O}\left(ND \log N\right)$ due to FFT. The core of WA is a linear-complexity variant of attention, and its complexity scales linearly with the sequence length as $\mathcal{O}\left(ND^2\right)$. Hence, the complexity of a SA is $\mathcal{O}\left(\max\left(ND^2, ND \log N\right)\right)$. As a result, the total complexity for the SAOT with L Transformer blocks is $\mathcal{O}\left(L \max\left(ND^2, ND \log N\right)\right)$.

Experiments

In this section, we first introduce the experimental setup. Next, we compare the performance between SAOT and more than ten baselines on six standard benchmarks to demonstrate the effectiveness of SAOT. Finally, we conduct the ablation study, make more comparisons, and evaluate the model's generalization ability.

Experimental Setup

Benchmarks Our experiments include six standard operator learning benchmarks: Darcy, Navier-Stokes (NS), Airfoil, Pipe, Plasticity, and Elasticity. The first two benchmarks are from FNO (Li et al. 2021), and the others are from geo-FNO (Li et al. 2023). These problems encompass various tasks in fluid and solid physical systems governed by different PDEs. Data for these problems is collected on uniform regular grids or irregular grid structures. Here we give a summary of these benchmarks and follow the experimental settings in the previous works (Li et al. 2021, 2023; Wu et al. 2023, 2024), which are summarized in Table 1. More detailed descriptions of these benchmarks can be found in the supplementary material.

Baselines We compare SAOT with various representative and powerful architectures as baselines: including **FNO** (Li et al. 2021) and its three variants: **U-NO** (Rahman, Ross, and Azizzadenesheli 2023), **Geo-FNO** (Li et al. 2023), and **F-FNO** (Tran et al. 2023); two Wavelet-based neural operators: **MWT** (Gupta, Xiao, and Bogdan 2021) and **WNO** (Tripura and Chakraborty 2022); and five advanced attention-based operators: **Galerkin** (Cao 2021), **OFormer** (Li, Meidani, and Farimani 2023), **GNOT** (Hao et al. 2023), **IPOT** (Lee and Oh 2024), and **Transolver** (Wu et al. 2024). Concretely, FNO and U-NO are only applicable to the Darcy and Navier-Stokes datasets since they are limited to regular grids. Geo-FNO extends the application of FNO to irregular grid datasets. MWT and WNO are two typical neural operators based on the Wavelet transform. The last five Transformer-based models design different attention mechanisms to handle complex geometries and enhance the performance. It's noteworthy that Transolver is the latest state-of-the-art model for operator learning.

Implementation For fair comparisons, all the experiments are conducted on one NVIDIA Tesla V100S-PCIE-32GB. Following the convention of related literature (Li et al. 2021, 2023; Wu et al. 2024), our model is trained with relative L^2 loss for 500 epochs with an initial learning rate of 1×10^{-3}. The relative L^2 error is also used as the evaluation metric to measure the quality of predictions.

Main Results

Table 2 reports the relative L^2 errors of our proposed SAOT and several baselines for six benchmarks. Except those of F-FNO, Galerkin, and Oformer cited from (Wu et al. 2024), the results of other baselines are cited from those original papers or reproduced by ourselves marked as "*". If the baseline cannot apply to a benchmark or the original papers did not report the results for this benchmark, we mark it as "/".

In Darcy and Navier-Stokes benchmarks, where the domains are discretized into uniform regular grids, SAOT achieves the lowest prediction errors, outperforming all the baselines. Specifically, SAOT significantly surpasses the spectral-based neural operators listed in Table 2 and also

Model	Darcy	NS	Airfoil	Pipe	Plasticity	Elasticity
FNO (2021)	0.0108	0.1556	/	/	/	/
U-NO* (2023)	0.0091	0.1408	/	/	/	/
Geo-FNO (2023)	0.0108	0.1556	0.0138	0.0067	0.0074	0.0229
F-FNO (2023)	0.0077	0.2322	0.0078	0.0070	0.0047	0.0263
MWT* (2021)	0.0067	0.1553	0.0076	0.0072	0.0027	0.0334
WNO* (2022)	0.0242	0.1613	0.0188	0.0070	/	0.0465
Galerkin (2021)	0.0084	0.1401	0.0118	0.0098	0.0120	0.0240
Oformer (2023)	0.0124	0.1705	0.0183	0.0168	0.0017	0.0183
GNOT (2023)	0.0105	0.1380	0.0076	/	/	0.0086
IPOT (2024)	0.0085	0.0885	0.0088	/	0.0033	0.0156
Transolver* (2024)	0.0058	0.0985	0.0053	**0.0043**	0.0012	**0.0067**
SAOT	**0.0049**	**0.0688**	**0.0048**	0.0063	**0.0008**	0.0080

Table 2: Performance comparisons with several baselines across six standard benchmarks. A smaller value indicates better performance. The best result is **bolded**, and the second best result is underlined. "*" means that the reported results are reproduced by us. "/" means that the result for this benchmark is not available.

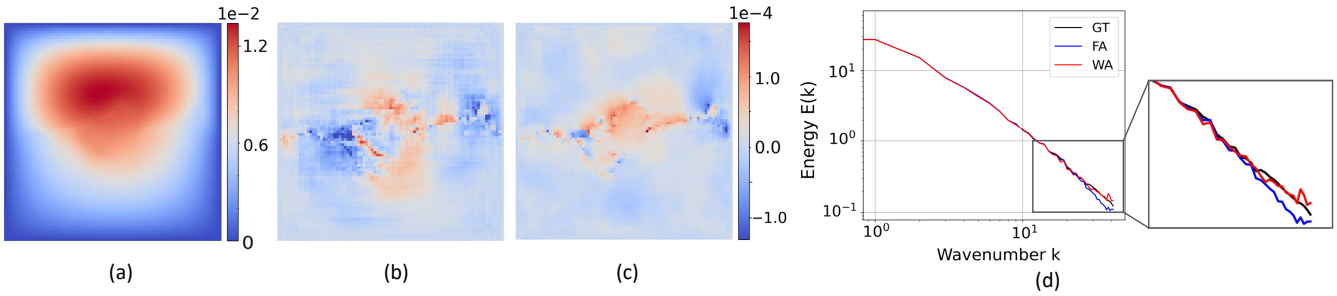

Figure 2: (a) Ground truth (GT); (b) Prediction errors of Fourier Attention (FA); (c) Prediction errors of Wavelet Attention (WA); (d) The energy spectrum of GT and the predictions of FA and WA with respect to the wavenumber. WA yields better-aligned energy spectrum predictions with true distributions than FA, especially at higher wavenumbers.

shows noticeable advantages over Transformer-based operators. In particular, compared to Transolver, SAOT gains remarkable improvement with relative promotions of 15.5% (0.0058→0.0049) in Darcy and 22.3% (0.0985→0.0688) in Navier-Stokes. In addition, SAOT exhibits excellent performance in Airfoil, Pipe, and Plasticity benchmarks. Especially, SAOT achieves remarkably lower errors than Transolver with relative performance promotions of 9.4% and 33.3% in Airfoil and Plasticity, respectively. The Elasticity benchmark, where the geometry is modeled using irregular point clouds, poses large challenges for operator learning. We can see that the above five spectral-based neural operators cannot or struggle to estimate the stress on unstructured point clouds, with errors exceeding 2%. In contrast, the latest models (GNOT/IPOT/Transolver) and our SAOT show the superiority of Transformer-based backbone models in handling complex geometries. Specifically, SAOT achieves the competitive performance with a relative L^2 error of 0.8% on this challenging benchmark. In addition to these quantitative comparisons, we also visualize the prediction results for an intuitive comparison in the supplementary material. Overall, SAOT exhibits its universal approximation capability across diverse datasets with complex geometries.

Ablation Study

To validate the capability of WA in capturing high-frequency details, we perform the experiments on the Darcy dataset and compare the performance of models using FA alone versus WA alone in Figure 2. We can see that the prediction errors of WA are relatively closer to 0 than those of FA. More importantly, Figure 2(d) shows that FA generates the predictions whose energy aligns well with the ground truth for only lower wavenumbers, but it ignores the higher frequencies, which can explain why FA often tends to generate over-smoothing solutions. Fortunately, our proposed WA mitigates this spectral bias by offering predictions with a better alignment on high-wavenumber modes.

To verify the effectiveness of mixing tokens in different frequency domains, we employ different attentions in SAOT and report the results in Table 3. We can see that WA achieves better performance than FA, especially for Elasticity with a remarkably lower prediction error from 0.0232→0.0129. As the subtle changes in a material structure can heavily impact the variations in stress distribution, the strength of WA to capture local details is beneficial for precise stress estimations and thus brings significant promotions. Furthermore, we find that WA shows a noticeably

Attn.	Darcy		Elasticity	
	Param. (M)	Rel. L2	Param. (M)	Rel. L2
FA	0.651	0.0058	0.576	0.0232
WA	2.361	0.0057	1.514	0.0129
SA	2.694	**0.0049**	2.040	**0.0080**

Table 3: Ablation study on the Darcy and Elasticity benchmarks about different attention mechanisms.

Model	Darcy		Elasticity	
	Param. (M)	Rel. L2	Param. (M)	Rel. L2
FNet	0.582	0.0168	0.466	0.1044
GFNet	8.067	0.0074	1.869	0.0230
AFNO	0.651	0.0058	0.576	0.0228
SAOT	2.694	**0.0049**	2.040	**0.0080**

Table 4: Performance comparisons between SAOT and other popular spectral Transformers in terms of the parameter count and relative L2 error.

better performance than the two Wavelet-based neural operators: MWT and WNO in Table 2, verifying the superiority of the proposed Wavelet-based attention. Ultimately, SA further significantly reduces the prediction errors on both benchmarks by combining the complementary features from FA and WA.

Comparison with Other Spectral Transformers

We further compare the proposed SAOT with three classical spectral Transformers: FNet (Lee-Thorp et al. 2021), GFNet (Rao et al. 2021), and AFNO (Guibas et al. 2022), which are proposed for vision tasks to perform token mixing in the Fourier domain. The results are shown in Table 4. FNet employs an unparameterized Fourier Transform to replace self-attention, leading to the smallest parameter size, but it results in the largest errors on two benchmarks. Owing to an element-wise multiplication between Fourier features and learnable global filters, GFNet experiences a dramatic rise in trainable network parameters, especially for large-scale grids in the Darcy dataset, yet its performance does not scale significantly with the increased parameters. AFNO performs better than FNet given similar parameter counts, but its error on the Elasticity dataset is still greater than 2%. Our SAOT model combines the advantages of two types of frequency-domain attentions under a modest parameter count, and thus achieves the lowest errors on both datasets.

Generalization Ability under Different Resolutions

To evaluate the model's generalization ability, we conduct the zero-shot super-resolution evaluation on the Darcy benchmark. Specifically, we train the model with the data on a resolution of 85^2 and then test the well-trained model with varying resolutions ranging from 43^2 to 421^2. Hence, we can assess the discretization-invariant property of our model

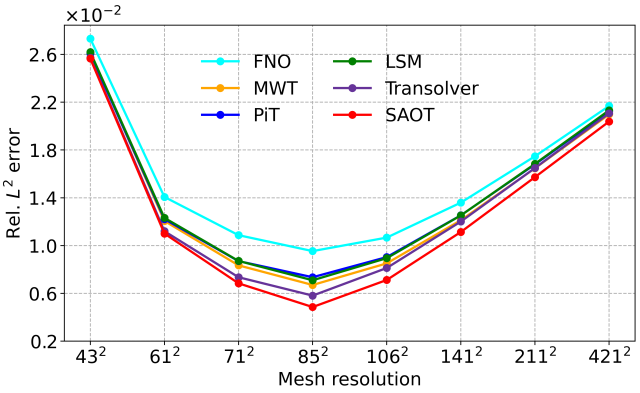

Figure 3: Model generalization performance on the Darcy dataset with different mesh resolutions.

and make comparisons with other methods: FNO (Li et al. 2021), MWT (Gupta, Xiao, and Bogdan 2021), LSM (Wu et al. 2023), PiT (Chen and Wu 2024), and Transolver (Wu et al. 2024). The comparison results are displayed in Figure 3. We can see that all the methods obtain the lowest errors when the test resolution matches the training resolution, i.e., 85^2. This is reasonable because the model is trained to be good at extracting features from data of this resolution. Inconsistency in the resolution between training and test data can introduce additional variability and complexity, making it harder for the model to make accurate predictions. Therefore, as the test resolution moves further away from 85^2, the test errors of all methods progressively rise, leading to U-shaped convex curves in Figure 3. Notably, our proposed SAOT model consistently yields lower errors across all test resolutions compared to the other competing methods, demonstrating its strong generalization capability and an excellent discretization-invariant property.

Conclusion

In this work, we propose the Spectral Attention Operator Transformer (SAOT), a novel framework for operator learning designed at the attention mechanism from a complementary frequency-domain perspective. Considering the difficulties of Fourier attention in capturing high-frequency components with local details, we first introduce a novel Wavelet attention with linear complexity, which harnesses the spatial-frequency localization of wavelets to efficiently extract locality-aware features and compensate for the spectral limitation of Fourier attention. We further introduce a spectral attention layer that combines a Wavelet attention and a parallel Fourier attention via a gated fusion block to leverage the complementary advantages of these two spectral operations. Experimental results demonstrate that the proposed Wavelet attention offers valuable high-frequency information to recover the local details that Fourier attention inherently lacks. Hence, our SAOT model can achieve superior performance over other state-of-the-art methods. This work reveals the Wavelet transform's potential in operator learning, opening new research avenues such as exploring alternative wavelet basis variants.

References

Azizzadenesheli, K.; Kovachki, N.; Li, Z.; Liu-Schiaffini, M.; Kossaifi, J.; and Anandkumar, A. 2024. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 1–9.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. arXiv:1607.06450.

Bachman, G.; Narici, L.; and Beckenstein, E. 2000. *Fourier and wavelet analysis*, volume 586. Springer.

Cao, S. 2021. Choose a Transformer: Fourier or Galerkin. In *NeurIPS*.

Cao, Y.; Fang, Z.; Wu, Y.; Zhou, D.-X.; and Gu, Q. 2021. Towards Understanding the Spectral Bias of Deep Learning. In *30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, 2205–2211. International Joint Conferences on Artificial Intelligence.

Chen, J.; and Wu, K. 2024. Positional Knowledge is All You Need: Position-induced Transformer (PiT) for Operator Learning. In *ICML*.

Cho, W.; Cho, S.; Jin, H.; Jeon, J.; Lee, K.; Hong, S.; Lee, D.; Choi, J.; and Park, N. 2024. Operator-learning-inspired modeling of neural ordinary differential equations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10): 11543–11551.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; Belanger, D. B.; Colwell, L. J.; and Weller, A. 2021. Rethinking Attention with Performers. In *ICLR*.

Clevert, D.-A. 2015. Fast and accurate deep network learning by exponential linear units (ELUs). arXiv:1511.07289.

Fanaskov, V.; and Oseledets, I. 2022. Spectral Neural Operators. arXiv:2205.10573.

Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2022. Adaptive fourier neural operators: Efficient token mixers for transformers. In *ICLR*.

Gupta, G.; Xiao, X.; and Bogdan, P. 2021. Multiwavelet-based Operator Learning for Differential Equations. In *NeurIPS*.

Hagnberger, J.; Kalimuthu, M.; Musekamp, D.; and Niepert, M. 2024. Vectorized Conditional Neural Fields: A Framework for Solving Time-dependent Parametric Partial Differential Equations. In *ICML*.

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. In *ICML*.

Hao, Z.; Ying, C.; Wang, Z.; Su, H.; Dong, Y.; Liu, S.; Cheng, Z.; Zhu, J.; and Song, J. 2023. GNOT: A General Neural Operator Transformer for Operator Learning. In *ICML*.

Jiang, P.; Yang, Z.; Wang, J.; Huang, C.; Xue, P.; Chakraborty, T.; Chen, X.; and Qian, Y. 2023. Efficient Super-Resolution of Near-Surface Climate Modeling Using the Fourier Neural Operator. *Journal of Advances in Modeling Earth Systems*, 15(7).

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020a. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *ICML*.

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020b. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*.

Khan, H.; and Yener, B. 2018. Learning filter widths of spectral decompositions with wavelets. *Advances in Neural Information Processing Systems*, 31.

Kitaev, N.; Kaiser, L.; and Levskaya, A. 2020. Reformer: The Efficient Transformer. In *ICLR*.

Kovachki, N.; Li, Z.; Liu, B.; Azizzadenesheli, K.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2023. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89): 1–97.

Lee, S.; and Oh, T. 2024. Inducing point operator transformer: A flexible and scalable architecture for solving pdes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1): 153–161.

Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontanon, S. 2021. Fnet: Mixing tokens with fourier transforms. arXiv:2105.03824.

Leng, J.; Ju, Y.; Duan, Y.; Zhang, J.; Lv, Q.; Wu, Z.; and Fan, H. 2025. FNIN: A Fourier Neural Operator-based Numerical Integration Network for Surface-from-gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5): 4580–4588.

Li, Y.; Du, T.; Pang, Y.; and Huang, Z. 2024. Component fourier neural operator for singularly perturbed differential equations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12): 13691–13699.

Li, Z.; Kovachki, N. B.; Azizzadenesheli, K.; liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2021. Fourier Neural Operator for Parametric Partial Differential Equations. In *ICLR*.

Li, Z.; Meidani, K.; and Farimani, A. B. 2023. Transformer for Partial Differential Equations' Operator Learning. *Transactions on Machine Learning Research*.

Li, Z.; Shu, D.; and Farimani, A. B. 2023. Scalable Transformer for PDE Surrogate Modeling. In *NeurIPS*.

Li, Z.-Y.; Huang, D. Z.; Liu, B.; and Anandkumar, A. 2023. Fourier Neural Operator with Learned Deformations for PDEs on General Geometries. *Journal of Machine Learning Research*, 24(388): 1–26.

Li, Z.-Y.; Kovachki, N. B.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2020. Neural Operator: Graph Kernel Network for Partial Differential Equations. arXiv:2003.03485.

Liu, L.; Liu, J.; Yuan, S.; Slabaugh, G.; Leonardis, A.; Zhou, W.; and Tian, Q. 2020. Wavelet-based dual-branch network for image demoiréing. In *ECCV*.

Liu, P.; Wang, P.; Ren, X.; Yuan, H.; Hao, Z.; Xu, C.; Cai, S.; and Ni, D. 2025. Aerogto: An efficient graph-transformer operator for learning large-scale aerodynamics of 3d vehicle

geometries. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 18924–18932.

Liu, X.; and Tang, H. 2024. DiffFNO: Diffusion Fourier Neural Operator. arXiv:2411.09911.

Liu-Schiaffini, M.; Berner, J.; Bonev, B.; Kurth, T.; Azizzadenesheli, K.; and Anandkumar, A. 2024. Neural operators with localized integral and differential kernels. In *ICML*.

Lu, L.; Jin, P.; and Karniadakis, G. E. 2019. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. arXiv:1910.03193.

Lu, L.; Jin, P.; Pang, G.; Zhang, Z.; and Karniadakis, G. E. 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3): 218–229.

Oommen, V.; Bora, A.; Zhang, Z.; and Karniadakis, G. E. 2025. Integrating neural operators with diffusion models improves spectral representation in turbulence modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 481(2309): 20240819.

Poli, M.; Massaroli, S.; Berto, F.; Park, J.; Dao, T.; Ré, C.; and Ermon, S. 2022. Transform once: Efficient operator learning in frequency domain. *Advances in Neural Information Processing Systems*, 35: 7947–7959.

Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.; Bengio, Y.; and Courville, A. 2019. On the spectral bias of neural networks. In *International conference on machine learning*, 5301–5310. PMLR.

Rahman, M. A.; Ross, Z. E.; and Azizzadenesheli, K. 2023. U-no: U-shaped neural operators. *Transactions on Machine Learning Research*.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global filter networks for image classification. In *NeurIPS*.

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

Roubíček, T. 2013. *Nonlinear partial differential equations with applications*. Springer Science & Business Media.

Tran, A.; Mathews, A.; Xie, L.; and Ong, C. S. 2023. Factorized Fourier Neural Operators. In *ICLR*.

Tripura, T.; and Chakraborty, S. 2022. Wavelet neural operator: a neural operator for parametric partial differential equations. arXiv:2205.02191.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*.

Wang, S.; Seidman, J. H.; Sankaran, S.; Wang, H.; Pappas, G. J.; and Perdikaris, P. 2024. Bridging Operator Learning and Conditioned Neural Fields: A Unifying Perspective. arXiv:2405.13998.

Wen, G.; Li, Z.; Azizzadenesheli, K.; Anandkumar, A.; and Benson, S. M. 2022. U-FNO–An enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*.

Wu, H.; Hu, T.; Luo, H.; Wang, J.; and Long, M. 2023. Solving High-Dimensional PDEs with Latent Spectral Models. In *ICML*.

Wu, H.; Luo, H.; Wang, H.; Wang, J.; and Long, M. 2024. Transolver: A Fast Transformer Solver for PDEs on General Geometries. In *ICML*.

Xiao, Z.; Hao, Z.; Lin, B.; Deng, Z.; and Su, H. 2024. Improved Operator Learning by Orthogonal Attention. In *ICML*.

Xiong, R.; Yang, Y.; He, D.; Zheng, K.; Zheng, S.; Xing, C.; Zhang, H.; Lan, Y.; Wang, L.; and Liu, T. 2020. On layer normalization in the transformer architecture. In *International conference on machine learning*, 10524–10533. PMLR.

Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; and Mei, T. 2022. Wavevit: Unifying wavelet and transformers for visual representation learning. In *ECCV*, 328–345. Springer.

Yue, X.; Zhu, L.; and Yang, Y. 2024. Point-Calibrated Spectral Neural Operators. arXiv:2410.11382.