# Spectral Convolutional Conditional Neural Process

**Peiman Mohseni**
Texas A&M University
peiman.mohseni@tamu.edu

**Nick Duffield**
Texas A&M University
duffieldng@tamu.edu

## Abstract

Neural processes (NPs) are probabilistic meta-learning models that map sets of observations to posterior predictive distributions, enabling inference at arbitrary domain points. Their capacity to handle variable-sized collections of unstructured observations, combined with simple maximum-likelihood training and uncertainty-aware predictions, makes them well-suited for modeling data over continuous domains. Since their introduction, several variants have been proposed. Early approaches typically represented observed data using finite-dimensional summary embeddings obtained through aggregation schemes such as mean pooling. However, this strategy fundamentally mismatches the infinite-dimensional nature of the generative processes that NPs aim to capture. Convolutional conditional neural processes (ConvCNPs) address this limitation by constructing infinite-dimensional functional embeddings processed through convolutional neural networks (CNNs) to enforce translation equivariance. Yet CNNs with local spatial kernels struggle to capture long-range dependencies without resorting to large kernels, which impose significant computational costs. To overcome this limitation, we propose the Spectral ConvCNP (SConvCNP), which performs global convolution in the frequency domain. Inspired by Fourier neural operators (FNOs) for learning solution operators of partial differential equations (PDEs), our approach directly parameterizes convolution kernels in the frequency domain, leveraging the relatively compact yet global Fourier representation of many natural signals. We validate the effectiveness of SConvCNP on both synthetic and real-world datasets, demonstrating how ideas from operator learning can advance the capabilities of NPs.

## 1 Introduction

Stochastic processes offer a mathematical framework for modeling systems that evolve with inherent randomness over continuous domains such as time and space. They underpin a wide range of scientific applications—from spatio-temporal climate dynamics to biological and physical systems—thereby motivating the development of machine learning methods that can learn from data generated by such stochastic phenomena [Mathieu et al., 2021, Vaughan et al., 2021, Allen et al., 2025, Ashman et al., 2025, Dupont et al., 2021]. Among classical approaches, Gaussian processes (GPs; Rasmussen et al., 2006) provide a Bayesian framework with closed-form inference and uncertainty quantification. However, their cubic computational cost from matrix inversion and the difficulty of specifying suitable kernels—especially in high-dimensional settings—limit their scalability.

Motivated by the success of deep neural networks in large-scale function approximation, neural network-based alternatives have emerged. Neural processes (NPs; Garnelo et al., 2018a,b) exemplify this paradigm, combining ideas from GPs and deep learning within a meta-learning framework. By exposing the model to multiple realizations of an underlying stochastic process, each treated as a distinct task, NPs learn shared structure across tasks to parameterize a neural mapping that *directly* approximates the corresponding posterior predictive distribution [Bruinsma, 2024]. Once trained, the model enables efficient probabilistic predictions on new tasks without further optimization.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Since the introduction of conditional neural processes (CNPs, Garnelo et al. [2018a]) as the first class within the NPs family [Jha et al., 2022], a wide range of extensions and variants have been proposed to enhance their effectiveness. One active line of research introduces explicit inductive biases into CNPs to encode the symmetries commonly present in scientific and physical domains [Gordon et al., 2019, Kawano et al., 2021, Holderrieth et al., 2021, Huang et al., 2023, Ashman et al., 2024a,b]. Another major direction seeks to move beyond the mean-field factorized Gaussian predictive distributions to which CNPs are limited. A popular approach augments CNPs with stochastic latent variables, giving rise to the family of latent neural processes (LNPs; Garnelo et al. [2018b], Louizos et al. [2019], Wang and Van Hoof [2020], Foong et al. [2020], Lee et al. [2020], Volpp et al. [2021], Wang et al. [2022], Wang and van Hoof [2022], Kim et al. [2022], Jung and Park [2023], Lee et al. [2023], Xu et al. [2023]). Alternative approaches explore autoregressive prediction schemes [Bruinsma et al., 2023, Nguyen and Grover, 2022], Gaussian predictive distributions with non-diagonal covariances [Bruinsma et al., 2021, Markou et al., 2022], and quantile-based parameterizations of the predictive distribution [Mohseni et al., 2023].

This work focuses on CNPs, particularly convolutional CNPs (ConvCNPs; Gordon et al. 2019), which were the first to endow NPs with translation equivariance. ConvCNPs introduce the convolutional deep set construction which characterizes a broad class of translation-equivariant mappings over finite, potentially unstructured sets of observations as a composition of functional embeddings with translation-equivariant operators, typically realized through convolutional neural networks (CNNs; Fukushima [1980], LeCun et al. [1989, 1998]).

Despite their effectiveness, ConvCNPs can struggle to aggregate information from observations spread across large spatial domains—a challenge that becomes particularly pronounced in sparse data regimes. This limitation stems from their reliance on local convolutional kernels with small receptive fields, which hampers their ability to model long-range dependencies. A natural remedy is to enlarge the kernel size to extend the receptive field; however, this approach rapidly increases the number of model parameters and computational cost [Romero et al., 2021]. Alternatively, transformer-based architectures can capture long-range interactions but incur quadratic, rather than linear, computational complexity in the number of observations [Vaswani et al., 2017, Nguyen and Grover, 2022].

In this work, we propose an alternative approach that represents functions in the frequency domain, inspired by the well-established observation that many natural processes exhibit energy concentration in low-frequency bands [Field, 1987, Ruderman and Bialek, 1993, Wainwright and Simoncelli, 1999]. This characteristic allows for efficient approximation using only a subset of dominant spectral coefficients, enabling tractable computations while preserving the signal's global structure. By parameterizing convolution kernels directly in the Fourier domain over a finite set of frequencies and leveraging the convolution theorem, our method attains large effective receptive fields without incurring prohibitive computational costs.

While spectral methods have been extensively studied in neural operator learning for partial differential equations (PDEs) (e.g., [Li et al., 2020a, Gupta et al., 2021, Kovachki et al., 2021, 2023, Helwig et al., 2023]), their application within NPs framework remains relatively unexplored. To bridge this gap, we propose the Spectral Convolutional Conditional Neural Process (SConvCNP)—a model that adopts Fourier neural operators (FNOs; Li et al. 2020a) to realize global convolution while maintaining computational efficiency. Across a suite of synthetic and real-world benchmarks, SConvCNP performs competitively with state-of-the-art baselines, illustrating how ideas from neural operators can enhance the flexibility and performance of NPs.

## 2 Preliminaries

### 2.1 Fourier Neural Operator

Neural operators [Chen and Chen, 1995, Li et al., 2020b,a, Kovachki et al., 2023, Raonic et al., 2023] constitute a class of neural networks that learn mappings between *function spaces* rather than finite-dimensional vectors. Analogous to conventional feed-forward architectures, they comprise stacked layers alternating between operator-based transformations and pointwise nonlinearities. Each transformation acts as an operator—typically a *linear* integral operator with kernel $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{Y}$—applied to an input function $v : \mathbb{X} \to \mathbb{Y}$:

$$\mathcal{K}[v](x) = \int \kappa(x, t)\, v(t)\, dt.$$

While this work focuses on linear operators, nonlinear formulations have also been explored, including continuous formulations of softmax attention. When the kernel is *stationary* ($\kappa(x,t) = \kappa(x-t)$), the operator simplifies to a convolution, $\mathcal{K}[v] = \kappa * v$, directly connecting neural operators to CNNs, where $\kappa$ is parameterized by learnable weights.

Typically, convolutional kernels are spatially local and possess a limited receptive field [Luo et al., 2016, Peng et al., 2017, Wang et al., 2018]. Consequently, modeling long-range dependencies requires increasingly large kernels, which substantially inflates the parameter count. The Fourier Neural Operator (FNO, Li et al. [2020a]) overcomes this limitation by exploiting the convolution theorem [Bracewell and Kahn, 1966, Oppenheim, 1999], which re-expresses convolution in the frequency domain as:

$$\mathcal{K}[v](x) = \mathcal{F}^{-1}\Big[\mathcal{F}[\kappa](\xi) \cdot \mathcal{F}[v](\xi)\Big](x), \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier transforms, respectively. Rather than parameterizing the kernel $\kappa$ in the spatial domain, the FNO learns its representation directly in the Fourier domain. When the input function $v$ is approximately band-limited—i.e., its spectrum $\mathcal{F}[v](\xi)$ carries negligible energy for frequencies $|\xi| > \xi_0$—the high-frequency components can be truncated with minimal loss of information. This property, observed in many natural signals [Field, 1987, Ruderman and Bialek, 1993, Wainwright and Simoncelli, 1999], allows setting $\mathcal{F}[\kappa](\xi) = 0$ outside the retained frequency band. In practice, the FNO parameterizes $\mathcal{F}[\kappa](\xi)$ only for a limited number of low-frequency modes, thereby capturing global dependencies while maintaining a compact parameterization.

In practice, functions are available only through discrete samples, requiring the use of the discrete Fourier transform (DFT) to transition between spatial and frequency domains. Given discretized samples of $v$ on a *uniform* grid $\mathcal{G} \subset \mathbb{X}$, the FNO computes their DFT, $\hat{\mathcal{F}}\{(v(x))_{x\in\mathcal{G}}\}$, efficiently using the fast Fourier transform (FFT; [Cooley and Tukey, 1965, Frigo and Johnson, 2005]). The resulting spectrum is truncated to retain only a finite subset of frequency modes, $\hat{\Xi} \subset \mathbb{R}^{d_{\mathbb{X}}}$, where $d_{\mathbb{X}}$ denotes the dimension of $\mathbb{X}$; these retained modes are assumed to capture most of the signal's energy. For each retained frequency mode $\hat{\xi} \in \hat{\Xi}$, the Fourier transform of the kernel, $\mathcal{F}[\kappa](\hat{\xi})$, is parameterized by learnable complex-valued weights. In the simplest case, this parameterization takes the form $\mathcal{F}[\kappa](\hat{\xi}) = w_{\hat{\xi}} \in \mathbb{C}^{d_{\mathbb{Y}}}$[1] where $d_{\mathbb{Y}}$ denotes the dimension of $\mathbb{Y}$. This formulation implicitly assumes periodicity in the spatial domain, as it models $\kappa$ using a discrete set of harmonics (i.e., a Dirac comb in frequency space). After pointwise multiplication in the frequency domain, the inverse FFT is applied to map the result back to the spatial domain, yielding the final operator output.

### 2.2 Neural Processes

Let $\mathfrak{P}$ denote the space of all $\mathbb{Y}$-valued stochastic processes on $\mathbb{X}$. Consider $\mathcal{P} \in \mathfrak{P}$, with $p(\cdot)$ denoting the density of its finite-dimensional distributions. A *task* $\mathcal{D}$ is defined as a finite collection of input–output pairs sampled from a realization $f \sim \mathcal{P}$, partitioned into a *context set* and a *target set*:

$$\mathcal{D} = (\mathcal{D}_c, \mathcal{D}_t) = \left(\{(x_{c,k}, y_{c,k})\}_{k\in\mathcal{I}_c},\ \{(x_{t,k}, y_{t,k})\}_{k\in\mathcal{I}_t}\right),$$

where

$$y_{c,k} = f(x_{c,k}) + \epsilon_{c,k}, \quad y_{t,k} = f(x_{t,k}) + \epsilon_{t,k}, \quad \epsilon_{c,k}, \epsilon_{t,k} \sim \mathcal{N}(0, \sigma^2_{\mathcal{P}}).$$

Here, $\sigma_{\mathcal{P}} > 0$ is the observation noise scale, and $\mathcal{I}_c, \mathcal{I}_t \subset \mathbb{N}$ are finite index sets corresponding to the context and target subsets, respectively. Neural processes (NPs; [Garnelo et al., 2018a,b]) constitute a class of models that employ neural networks to learn a mapping $\eta : \bigcup_{k=0}^{\infty}(\mathbb{X}\times\mathbb{Y})^k \to \mathfrak{P}$, which takes a finite context set $\mathcal{D}_c$ and gives a *direct* approximation to the posterior predictive process [Bruinsma et al., 2021, Bruinsma, 2024, Ashman et al., 2024a,b]. Let $q_{\mathcal{D}_c}(\cdot)$ denote the density of $\eta[\mathcal{D}_c]\big((x_{t,k})_{k\in\mathcal{I}_t}\big)$. The NP approximation can then be expressed as

$$q_{\mathcal{D}_c}((y_{t,k})_{k\in\mathcal{I}_t} \mid (x_{t,k})_{k\in\mathcal{I}_t}, \mathcal{D}_c) \approx p((y_{t,k})_{k\in\mathcal{I}_t} \mid (x_{t,k})_{k\in\mathcal{I}_t}, \mathcal{D}_c)\,.$$

In this work, we focus on conditional NPs (CNPs, Garnelo et al. [2018a]), which restrict $q_{\mathcal{D}_c}$ to the family of mean-field Gaussians, i.e. $q_{\mathcal{D}_c}((y_{t,k})_{k\in\mathcal{I}_t} \mid (x_{t,k})_{k\in\mathcal{I}_t}, \mathcal{D}_c) = \prod_{k\in\mathcal{I}_t} q_{\mathcal{D}_c}(y_{t,k} \mid x_{t,k}, \mathcal{D}_c)$.

[1] More generally, a matrix-valued parameterization $\mathcal{F}[\kappa](\hat{\xi}) = w_{\hat{\xi}} \in \mathbb{C}^{c_{\text{out}}\times d_{\mathbb{Y}}}$ is used, where $c_{\text{out}}$ is the number of output channels, allowing for joint mixing of input channels and projection into a space with potentially different dimensionality.

In general, CNPs parameterization of the predictive distribution $q_{\mathcal{D}_c}(y_{t,k} \mid x_{t,k}, \mathcal{D}_c)$ can be abstracted as a two-stage encoder-decoder pipeline [Bruinsma, 2024, Ashman et al., 2024a, 2025]. The encoder $\varphi_e : \mathbb{X} \times \bigcup_{k=0}^{\infty}(\mathbb{X} \times \mathbb{Y})^k \to \mathbb{H}$ maps the target input and context set into a latent representation, while the decoder $\varphi_d : \mathbb{X} \times \mathbb{H} \to \Theta$ transforms the target input and latent code into parameters $\theta \in \Theta$ of the predictive distribution $q_{\mathcal{D}_c}(y_{t,k} \mid x_{t,k}, \mathcal{D}_c)$.

The vanilla CNP encodes context pairs $(x_{c,k}, y_{c,k})$ using a *permutation-invariant* scheme [Qi et al., 2017, Zaheer et al., 2017]: each pair is independently mapped to a representation $\varepsilon_k$, which are then aggregated—typically by averaging—into a single embedding $\varepsilon \in \mathbb{H}$, where $\mathbb{H}$ is a *finite-dimensional* Euclidean space. The target input does not affect this encoding (i.e., $\varphi_e$ is constant with respect to it). The decoder then combines the target input with $\varepsilon$ through a feedforward network to produce the parameters of the Gaussian predictive distribution.

Although sum-pooling aggregation provides universal approximation guarantees [Zaheer et al., 2017, Bloem-Reddy and Teh, 2020], NPs employing such mechanisms often exhibit underfitting in practice [Kim et al., 2019]. Prior works have partly attributed this phenomena to two primary factors[Xu et al., 2020]: (1) the limitation of summaries with *prespecified finite dimensionality* in representing context sets of arbitrary size [Wagstaff et al., 2019], and (2) the shortcomings of simple sum or mean pooling operations to effectively capture rich interactions between context and target points[Xu et al., 2020, Nguyen and Grover, 2022].

Given that the problems tackled by NPs are inherently *functional* in nature, it is natural to seek embeddings that also exhibit a functional character. In this spirit, Gordon et al. [2019] formalize a general framework for constructing translation-equivariant prediction maps over sets, satisfying

$$\eta\Big[\big\{(x+\tau, y) \mid (x,y) \in \mathcal{D}_c\big\}\Big]\big((x_{t,k})_{k\in\mathcal{I}_t}\big) \;=\; \eta\big[\mathcal{D}_c\big]\big((x_{t,k}-\tau)_{k\in\mathcal{I}_t}\big) \quad \forall\, \tau \in \mathbb{X}.$$

They show that a broad family of such maps can be expressed as $\eta[\mathcal{D}_c] = \varphi_d\big[\varphi_e\left[\mathcal{D}_c\right]\big]$, where the functional embedding is defined by

$$\varphi_e[\mathcal{D}_c](\cdot) = \sum_{(x_k^{(c)}, y_k^{(c)}) \in \mathcal{D}_c} \phi(y_k^{(c)})\, \psi_e(\cdot - x_k^{(c)}), \tag{2}$$

and $\varphi_d : \mathbb{H} \to C_b(\mathbb{X}, \mathbb{Y})$ is a translation-equivariant decoder operating on a function space $\mathbb{H}$. Here, $C_b(\mathbb{X}, \mathbb{Y})$ denotes the space of bounded continuous functions from $\mathbb{X}$ to $\mathbb{Y}$; $\phi(y) = (1, y)$[2] and $\psi_e : \mathbb{X} \to \mathbb{R}$ is a continuous, strictly positive-definite kernel typically chosen to be a Gaussian.

In implementing the ConvCNP, the encoder output $\varphi_e[\mathcal{D}_c]$ is first evaluated on a *uniform* grid $\mathcal{G} \subset \mathbb{X}$ that spans the *joint* support of both the context and target points. This yields a discretized representation $(\varphi_e[\mathcal{D}_c](x))_{x\in\mathcal{G}}$, which is then passed through the decoder $\varphi_d$ to produce $\big(\varphi_d\big[\varphi_e[\mathcal{D}_c]\big](x)\big)_{x\in\mathcal{G}}$. Because the target inputs $(x_{t,k})_{k\in\mathcal{I}_t}$ may not lie exactly on the grid $\mathcal{G}$, ConvCNP employs an interpolation step to obtain predictions at arbitrary target locations. Specifically

$$(\theta_{t,k})_{k\in\mathcal{I}_t} = \left(\sum_{x\in\mathcal{G}} \varphi_d\big[\varphi_e[\mathcal{D}_c]\big](x)\, \psi_d(x_{t,k} - x)\right)_{k\in\mathcal{I}_t}, \tag{3}$$

where $\psi_d$ is another strictly positive kernel. Note that this step can be viewed as part of the decoder itself, thereby preserving the overall encoder–decoder abstraction discussed earlier.

## 3 Spectral Convolutional Conditional Neural Process

The decoder $\varphi_d$ in ConvCNP is typically parameterized using standard CNNs such as U-Net [Ronneberger et al., 2015] or ResNet [He et al., 2016]. These architectures employ discrete convolutional kernels—finite sets of learnable parameters that define localized filters operating over neighboring grid points. The kernel size, fixed *a priori*, determines the receptive field of each convolution [Ding et al., 2022] and is generally much smaller than the overall spatial extent of the input signals [Romero et al., 2021, Knigge et al., 2023].

[2]For most applications, this representation suffices. Generally, $\phi(y) = (1, y, \ldots, y^M)$, where $M$ is the multiplicity of repeated inputs. See Gordon et al. [2019] and Bruinsma [2024] for further discussion.

This locality constraint limits the model's capacity to capture long-range dependencies and to integrate information from observations distributed across wide spatial domains [Peng et al., 2017, Wang et al., 2018, Ramachandran et al., 2019, Wang et al., 2020]. The issue becomes particularly pronounced when handling sparse or irregularly sampled data, where effective global reasoning cannot emerge solely from local convolution operations. While increasing the kernel size could enlarge the receptive field, it leads to a rapid growth in both parameter count and computational cost. Transformer-based architectures offer an alternative by enabling global interactions [Vaswani et al., 2017]; however, they typically incur quadratic rather than linear complexity in the number of input locations [Nguyen and Grover, 2022, Feng et al., 2022, Ashman et al., 2024a, 2025].

To overcome this limitation without relying on prohibitively large filters or computationally expensive transformers, we exploit the Fourier representation of natural signals. This choice is motivated by the well-established observation that many natural signals are approximately band-limited (see Section 2.1), implying that their Fourier representation offers a more compact encoding while preserving the global structure of the signal compared to its spatial-domain representation. Specifically, we realize the operator $\varphi_d$ via spectral convolution modules based on equation 1 [Li et al., 2020a]. This substitution enables global convolution that captures long-range dependencies from sparse or irregularly sampled data—without causing a parameter explosion. We refer to the resulting model as the Spectral Convolutional Conditional Neural Process (SConvCNP).

**Computational Complexity.** The computational cost of the SConvCNP comprises three parts: (1) $\mathcal{O}(|\mathcal{D}_c||\mathcal{G}|)$ for discretizing the functional embedding on grid $\mathcal{G}$ (equation 2); (2) $\mathcal{O}(|\mathcal{G}|\log|\mathcal{G}|)$ for spectral convolution via FFT (equation 1); and (3) $\mathcal{O}(|\mathcal{D}_t||\mathcal{G}|)$ for interpolating outputs at target locations (equation 3). Overall, the complexity is $\mathcal{O}(|\mathcal{G}|(|\mathcal{D}_c| + \log|\mathcal{G}| + |\mathcal{D}_t|))$, comparable to ConvCNP's $\mathcal{O}(|\mathcal{G}|(|\mathcal{D}_c| + 1 + |\mathcal{D}_t|))$, both scaling *linearly* with the task size. In contrast, transformer NPs (TNPs, Kim et al. [2019], Nguyen and Grover [2022], Feng et al. [2022], Ashman et al. [2024a]) scale quadratically, $\mathcal{O}(|\mathcal{D}_c|^2 + |\mathcal{D}_c||\mathcal{D}_t|)$. Thus, SConvCNP and ConvCNP are more efficient for large datasets but limited by the exponential growth of $|\mathcal{G}|$ with input dimensionality. TNPs, while more expensive in the task size, handle high-dimensional inputs more effectively where grid-based methods become infeasible.

**Positional Encodings** The convolution operator preserves translation equivariance under the Fourier transform (see Equation 1). However, practical FNOs implementations often include explicit positional information to improve predictive accuracy [Li et al., 2020a, Tran et al., 2021, Gupta et al., 2021, Rahman et al., 2022c, Helwig et al., 2023, Tripura and Chakraborty, 2023, Liu et al., Li et al., 2024]. Accordingly, we augment the functional embedding $\varphi_e[\mathcal{D}_c](x)$ with absolute positional features:

$$\widetilde{\varphi}_e[\mathcal{D}_c](x) = \big(\varphi_e[\mathcal{D}_c](x),\, x\big).$$

Although this addition breaks translation equivariance, it consistently improves performance (see Section C.4 for ablation results), aligning with prior observations. Future work may explore *relative* positional encodings [Shaw et al., 2018, Su et al., 2024], which preserve translation equivariance while providing spatial context.

**Discretization Sensitivity of DFT.** Unlike the continuous Fourier transform, DFT—and by extension FFT—is inherently sensitive to the grid $\mathcal{G}$ on which $\varphi_e[\mathcal{D}_c]$ (or $\widetilde{\varphi}_e[\mathcal{D}_c]$) is discretized. This sensitivity stems from the discretization resolution and the spatial range of the domain; varying either leads to mismatched Fourier representations and altered behavior in spectral convolution modules (see Section A). Resolution dependence is not unique to the DFT: CNN-based operator learning models also exhibit this behavior [Raonic et al., 2023, Bartolucci et al., 2023]. ConvCNP, in particular, mitigates it by fixing the grid resolution. While spatial CNNs remain stable once resolution is fixed, the DFT is sensitive unless *both* resolution and spatial range are controlled. Accordingly, we fix both parameters, choosing a range large enough to cover all context and target inputs across tasks. When this is impractical, a patch-based strategy can be used—dividing the domain into (possibly overlapping) fixed-size patches, applying spectral convolutions independently, and aggregating the outputs. This parallels standard convolutional modules but allows much larger receptive fields. Similar ideas in transformer models for reducing computational cost suggest a promising direction for future work [Beltagy et al., 2020, Zaheer et al., 2020, Liu et al., 2021, Ding et al., 2023].

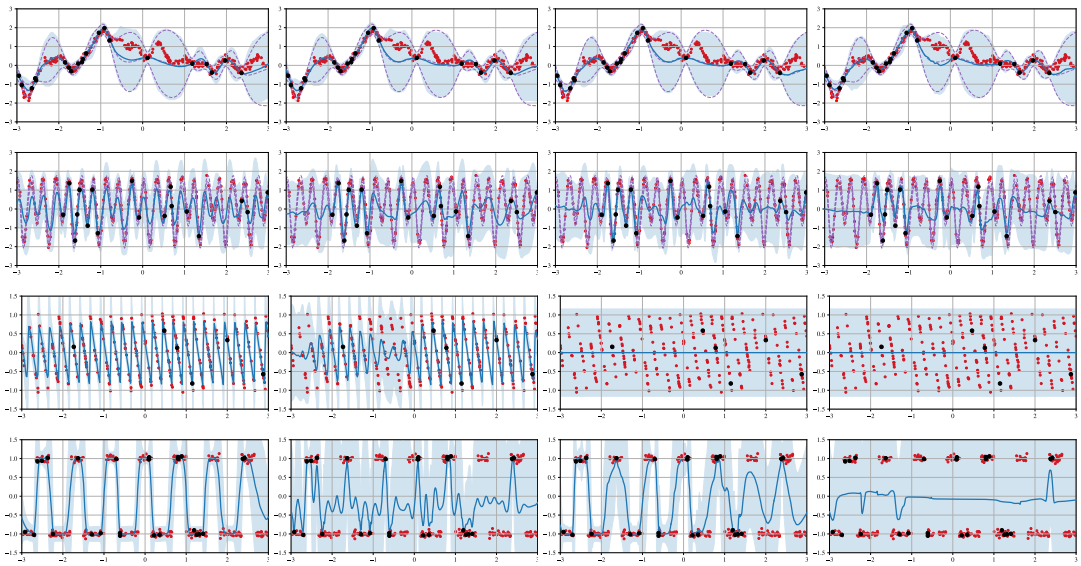


Figure 1: Examples of predictions made by different methods on synthetic datasets. Rows correspond to different types of underlying functions (Matérn 5/2 GP, Periodic GP, Sawtooth, and Square Wave), while columns represent the methods being compared: SConvCNP (our method), ConvCNP, TETNP, and TNP. Each plot shows context points (black), target points (red), and the model predictions with uncertainty (blue lines and shaded areas).

## 4 Experiments

We evaluate our framework on three standard regression benchmarks and compare its performance against several representative models from the Conditional Neural Processes (CNP) family. Specifically, we include the original CNP [Garnelo et al., 2018a], the Attentive CNP (ACNP; Kim et al. 2019), the Convolutional CNP (ConvCNP; Gordon et al. 2019), the Transformer Neural Process (TNP; Nguyen and Grover 2022), and the Translation-Equivariant Transformer Neural Process (TETNP; Ashman et al. 2024a). To evaluate model performance, we report both the log-likelihood and root-mean-squared error (RMSE) metrics. The implementation and experimental code are available at https://github.com/peiman-m/SConvCNP.

### 4.1 Synthetic 1-D Regression

We train our model on mini-batches of 16 tasks. Each epoch processes 1,000 batches, and we train for 250 epochs—exposing the model to 4 million tasks in total. We generate tasks using four distinct stochastic processes: two Gaussian processes (GPs) with periodic and Matérn 5/2 kernels, and two non-Gaussian signals—sawtooth and square waves—with randomly sampled hyperparameters. For each task, the number of context and target points is drawn independently as $n_c \sim \mathcal{U}[5, 25)$ and $n_t \sim \mathcal{U}[5, 25)$, and input locations are sampled uniformly from $[-3, 3)$. Validation is performed on a fixed meta-dataset of 4,096 tasks, with evaluation after every epoch. Final performance is measured on a held-out meta-dataset of 64,000 tasks. Note that while training tasks are generated dynamically, validation and test sets are fixed. Table 1 reports predictive performance across all methods. SConvCNP consistently matches or outperforms baselines, with particularly strong gains on periodic functions—suggesting that global convolutional structure enables more effective modeling of periodic patterns. Figure 1 illustrates several representative cases where SConvCNP yields superior fits. Detailed experimental configurations are provided in Appendix.

Table 1: Comparison of predictive performance (RMSE and log-likelihood) obtained by different methods over synthetically generated tasks (4 Seeds). Lower RMSE and higher log-likelihood values indicate better performance.

| Metric | Data | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | CNP | AttCNP | TNP | TETNP | ConvCNP | SConvCNP |
| RMSE ↓ | Matérn 5/2 | $0.49_{\pm 0.00}$ | $0.45_{\pm 0.00}$ | $0.44_{\pm 0.00}$ | $0.44_{\pm 0.00}$ | $0.44_{\pm 0.00}$ | $0.44_{\pm 0.00}$ |
| | Periodic | $0.81_{\pm 0.00}$ | $0.67_{\pm 0.03}$ | $0.61_{\pm 0.00}$ | $0.51_{\pm 0.01}$ | $0.61_{\pm 0.03}$ | $0.48_{\pm 0.00}$ |
| | Sawtooth | $0.57_{\pm 0.00}$ | $0.57_{\pm 0.00}$ | $0.57_{\pm 0.00}$ | $0.57_{\pm 0.00}$ | $0.36_{\pm 0.07}$ | $0.19_{\pm 0.00}$ |
| | Square Wave | $0.97_{\pm 0.00}$ | $0.97_{\pm 0.00}$ | $0.73_{\pm 0.00}$ | $0.83_{\pm 0.01}$ | $0.90_{\pm 0.00}$ | $0.73_{\pm 0.00}$ |
| LogLike ↑ | Matérn 5/2 | $-0.54_{\pm 0.00}$ | $-0.31_{\pm 0.00}$ | $-0.29_{\pm 0.00}$ | $-0.27_{\pm 0.00}$ | $-0.29_{\pm 0.00}$ | $-0.29_{\pm 0.00}$ |
| | Periodic | $-1.19_{\pm 0.00}$ | $-0.93_{\pm 0.08}$ | $-0.77_{\pm 0.01}$ | $-0.60_{\pm 0.03}$ | $-0.79_{\pm 0.05}$ | $-0.56_{\pm 0.00}$ |
| | Sawtooth | $-0.87_{\pm 0.00}$ | $-0.87_{\pm 0.00}$ | $-0.87_{\pm 0.00}$ | $-0.87_{\pm 0.00}$ | $0.19_{\pm 0.37}$ | $1.04_{\pm 0.02}$ |
| | Square Wave | $-1.39_{\pm 0.00}$ | $-1.39_{\pm 0.00}$ | $-0.86_{\pm 0.00}$ | $-1.12_{\pm 0.01}$ | $-1.25_{\pm 0.01}$ | $-0.86_{\pm 0.00}$ |

### 4.2 Predator–Prey Model

We next evaluate performance on trajectories sampled from a stochastic version [Bruinsma et al., 2023] of the Lotka–Volterra equations [Lotka, 1910, Volterra, 1926]:

$$dX_t = \alpha X_t\,dt - \beta Y_t X_t\,dt + \sigma X_t^{\nu}\,dW_t^{(1)}, \quad dY_t = -\gamma Y_t\,dt + \delta Y_t X_t\,dt + \sigma Y_t^{\nu}\,dW_t^{(2)}. \tag{4}$$

Here, $X_t$ and $Y_t$ denote prey and predator populations, respectively. Prey grow exponentially at rate $\alpha$, while predators decline at rate $\gamma$, with interaction terms $\beta$ and $\delta$ modeling consumption and reproductive gains. In the deterministic core, larger $\alpha$ or $\delta$ tend to increase the amplitude and frequency of natural oscillations, whereas higher $\beta$ or $\gamma$ act to damp cycles and shorten periods. Stochasticity is introduced via independent Brownian motions $W_t^{(1)}$ and $W_t^{(2)}$, with magnitude $\sigma$ controlling overall noise intensity and exponent $\nu$ governing how fluctuations scale with population size (e.g. linear for $\nu = 1$, super- or sub-linear otherwise). To construct the meta-dataset, we simulate these equations on a dense time grid spanning 110 years, discarding the first 10 years as burn-in. Each task is formed by sampling $n_c + n_t$ input–output pairs $(t, (X_t, Y_t))$ from the trajectories, where $n_c \sim \mathcal{U}[5, 25)$ and $n_t \sim \mathcal{U}[5, 25)$. The input times $t$ are sampled uniformly from the post-burn-in interval. For model training and inference, we scale the time values by a factor of $0.1$, mapping the 100-year period to the range $[0, 10]$. Similarly, population values $X_t$ and $Y_t$ are scaled by $0.01$ to improve numerical stability and model convergence. As with the previous experiment, training tasks are generated on-the-fly, while validation and test sets are held fixed.

Table 2: Comparison of predictive performance (RMSE ↓ and Log-likelihood ↑) obtained by different methods on the Lotka-Volterra predator-prey simulation (4 Seeds).

| | CNP | AttCNP | TNP | TETNP | ConvCNP | SConvCNP |
|---|---|---|---|---|---|---|
| RMSE ↓ | $1.47_{\pm 0.00}$ | $1.29_{\pm 0.01}$ | $1.23_{\pm 0.00}$ | $1.16_{\pm 0.00}$ | $1.19_{\pm 0.00}$ | $1.19_{\pm 0.00}$ |
| Log-likelihood ↑ | $-1.66_{\pm 0.00}$ | $-1.27_{\pm 0.05}$ | $-1.14_{\pm 0.00}$ | $-0.93_{\pm 0.00}$ | $-1.01_{\pm 0.00}$ | $-1.01_{\pm 0.00}$ |

### 4.3 Traffic Flow

For our final experiment, we utilized the California traffic flow dataset from LargeST [Liu et al., 2023b], a large-scale benchmark for traffic forecasting. This dataset encompasses traffic flow readings from 8,600 loop detector sensors installed across California's highway system, with a temporal coverage spanning 5 years (2017-2021) sampled at 5-minute intervals. We choose to focus on the year 2020, which includes significant anomalies due to the COVID-19 pandemic, and is expected to exhibit greater variance in traffic patterns. As a preprocessing step, we discard any sensor with more than 50% missing data. We randomly partition the remaining sensors into training, validation, and test sets using a 6:1:3 ratio. Each sensor's year-long data is segmented into non-overlapping

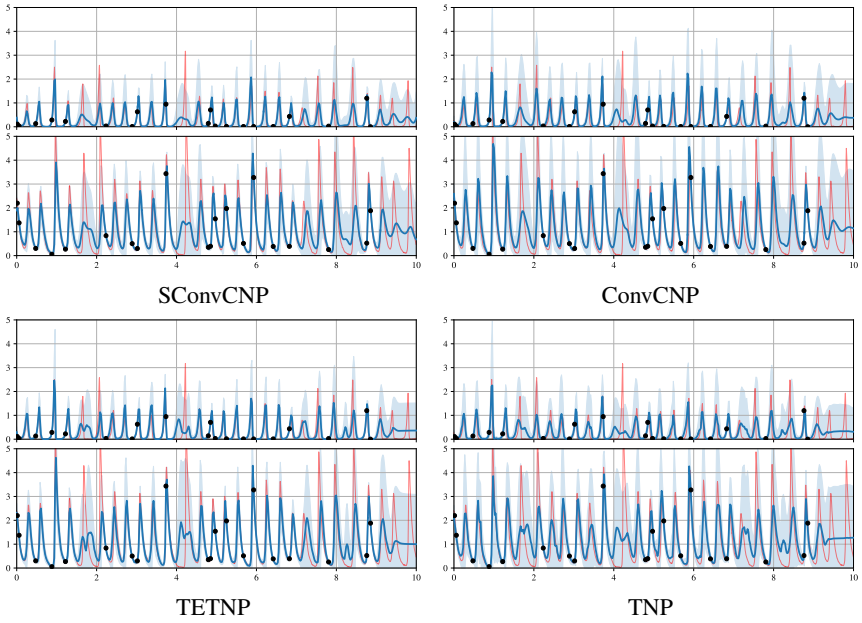


Figure 2: Examples of predictions for the Lotka-Volterra predator-prey dynamics by different methods. Each panel shows context points (black), target points (red), and model predictions with uncertainty bounds (blue).

14-day patches. Each patch forms a dense trajectory from which we sample context and target points, similar to our previous experiments. This design enables the model to handle both short- and long-range dependencies under realistic, non-stationary conditions. Evaluation is conducted using log-likelihood and RMSE metrics, consistent with prior benchmarks. Our results demonstrate that Spectral-ConvCNP remains robust under highly variable real-world data and outperforms baseline methods on both accuracy and uncertainty estimation.

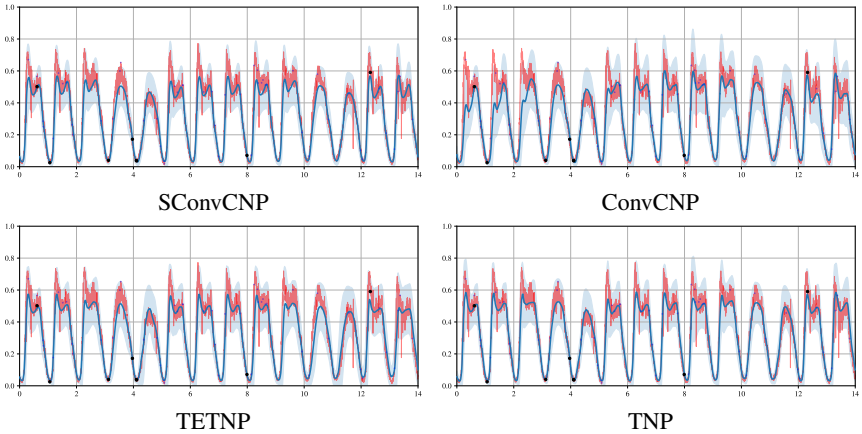


Figure 3: Examples of predictions on the California traffic flow dataset by different methods. Each panel shows context points (black), target points (red), and model predictions with uncertainty (blue). SConvCNP effectively captures both short-term patterns and long-range dependencies in traffic flow data compared to ConvCNP, TETNP, and TNP.

Table 3: Comparison of predictive performance (RMSE ↓ and Log-likelihood ↑) obtained by different methods on the California traffic flow dataset (4 Seeds).

| | CNP | AttCNP | TNP | TETNP | ConvCNP | SConvCNP |
|---|---|---|---|---|---|---|
| RMSE ↓ | $0.04_{\pm 0.00}$ | $0.04_{\pm 0.00}$ | $0.05_{\pm 0.00}$ | $0.04_{\pm 0.00}$ | $0.04_{\pm 0.00}$ | $0.03_{\pm 0.00}$ |
| Log-likelihood ↑ | $1.88_{\pm 0.00}$ | $1.84_{\pm 0.00}$ | $2.02_{\pm 0.01}$ | $1.79_{\pm 0.01}$ | $1.97_{\pm 0.00}$ | $2.07_{\pm 0.00}$ |

# 5 Related Works

## 5.1 Neural PDE Solvers

The substantial computational demands posed by conventional numerical solvers for partial differential equations (PDEs), which are commonly encountered in mathematically oriented scientific domains, have spurred a growing interest in employing machine learning towards improved and computationally efficient alternatives Gupta and Brandstetter [2022]. Among these potential alternatives, neural operators [Li et al., 2020b, Kovachki et al., 2023], and specifically Fourier neural operators (FNOs, Li et al. [2020a]), have risen as a particularly successful and promising approach. Since then, various enhancements have been introduced. For instance, Helwig et al. [2023] expanded group convolutions into the frequency domain, developing Fourier layers that maintain equivariance with respect to rotations, translations, and reflections. Gupta et al. [2021] introduced multiwavelet-based neural operators, achieved by parameterizing the kernel's projection onto predefined multiwavelet polynomial bases. Tran et al. [2021] use a separable Fourier representation along with enhanced residual connections to decrease model complexity and allow for deeper architectures. In another work, Rahman et al. [2022a] introduced a generative model framework to learn distributions over function spaces through the use of neural operators. Liu et al. [2023a] proposed an integral neural operator architecture designed to exhibit both translation and rotation invariance.

## 5.2 Function Space Inference

In non-parametric Bayesian modeling, GPs and deep GPs [Damianou and Lawrence, 2013] exemplify function-space priors that offer uncertainty estimates but are computationally infeasible for large datasets. This limitation has motivated alternatives such as Bayesian neural networks (BNNs; Hinton and Van Camp [1993], Neal [2012]), which combine the scalability of neural networks with Bayesian uncertainty. However, defining meaningful priors over network weights remains challenging. Recent efforts reformulate Bayesian inference in neural networks as inferring a posterior over functions induced by the weights. Variational implicit processes (VIPs; Ma et al. [2019], Santana et al. [2021], Ortega et al. [2022]) generalize GPs through implicit distributions over random variables, while functional variational BNNs (fBNNs; Sun et al. [2019]) align BNNs with priors by minimizing functional KL divergence, though this approach faces issues of intractability and well-definedness [Burt et al., 2020]. Follow-up work [Ma and Hernández-Lobato, 2021, Rudner et al., 2022, Wild et al., 2022] addresses these limitations. Parallel research on neural processes (NPs; Garnelo et al. [2018a,b]) uses neural networks to parameterize stochastic processes, with extensions to temporal settings [Singh et al., 2019, Yoon et al., 2020]. More recently, Dutordoir et al. [2023] and Mathieu et al. [2023] extend diffusion models to stochastic processes via their finite marginals.

# 6 Conclusion

In this work, we introduced the Spectral Convolutional Conditional Neural Process (SConvCNP), a new addition to the CNPs family that harnesses advancements in operator learning to enhance the expressive capabilities of the Convolutional Conditional Neural Process (ConvCNP) when modeling stochastic processes. Our experiments, conducted on synthetic datasets, demonstrated that SConvCNP enhances the predictive performance of ConvCNP in regression tasks, as evidenced by improvements in log-likelihood. Furthermore, they adeptly capture global symmetries, including the prevalent periodic patterns in the data.

## Acknowledgments

We thank Arman Hasanzadeh and Jonathan W. Siegel for their insightful feedback and constructive suggestions. We also acknowledge the Texas A&M High Performance Research Computing facility for providing the computational resources used in this study. Finally, we thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this work.

## References

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, 2025.

Matthew Ashman, Cristiana Diaconu, Junhyuck Kim, Lakee Sivaraya, Stratis Markou, James Requeima, Wessel P Bruinsma, and Richard E Turner. Translation equivariant transformer neural processes. *arXiv preprint arXiv:2406.12409*, 2024a.

Matthew Ashman, Cristiana Diaconu, Adrian Weller, Wessel Bruinsma, and Richard Turner. Approximately equivariant neural processes. *Advances in Neural Information Processing Systems*, 37: 97088–97123, 2024b.

Matthew Ashman, Cristiana Diaconu, Eric Langezaal, Adrian Weller, and Richard E. Turner. Gridded transformer neural processes for spatio-temporal data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=O0oe7hPtbl.

Francesca Bartolucci, Emmanuel de Bezenac, Bogdan Raonic, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. *Advances in Neural Information Processing Systems*, 36:69661–69672, 2023.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020.

Ron Bracewell and Peter B Kahn. The fourier transform and its applications. *American Journal of Physics*, 34(8):712–712, 1966.

Wessel P Bruinsma. Convolutional conditional neural processes. *arXiv preprint arXiv:2408.09583*, 2024.

Wessel P Bruinsma, James Requeima, Andrew YK Foong, Jonathan Gordon, and Richard E Turner. The gaussian neural process. *arXiv preprint arXiv:2101.03606*, 2021.

Wessel P Bruinsma, Stratis Markou, James Requiema, Andrew YK Foong, Tom R Andersson, Anna Vaughan, Anthony Buonomo, J Scott Hosking, and Richard E Turner. Autoregressive conditional neural processes. *arXiv preprint arXiv:2303.14468*, 2023.

David R Burt, Sebastian W Ober, Adrià Garriga-Alonso, and Mark van der Wilk. Understanding variational inference in function-space. *arXiv preprint arXiv:2011.09421*, 2020.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE transactions on neural networks*, 6(4):911–917, 1995.

James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR, 2013.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022.

Emilien Dupont, Yee Whye Teh, and Arnaud Doucet. Generative models as distributions of functions. *arXiv preprint arXiv:2102.04776*, 2021.

Vincent Dutordoir, Alan Saul, Zoubin Ghahramani, and Fergus Simpson. Neural diffusion processes. In *International Conference on Machine Learning*, pages 8990–9012. PMLR, 2023.

Leo Feng, Hossein Hajimirsadeghi, Yoshua Bengio, and Mohamed Osama Ahmed. Latent bottlenecked attentive neural processes. *arXiv preprint arXiv:2211.08458*, 2022.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. *Advances in Neural Information Processing Systems*, 33:8284–8295, 2020.

Matteo Frigo and Steven G Johnson. The design and implementation of fftw3. *Proceedings of the IEEE*, 93(2):216–231, 2005.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pages 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Jonathan Gordon, Wessel P Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. *arXiv preprint arXiv:1910.13556*, 2019.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. *Advances in neural information processing systems*, 34:24048–24062, 2021.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

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*, pages 770–778, 2016.

Jacob Helwig, Xuan Zhang, Cong Fu, Jerry Kurtin, Stephan Wojtowytsch, and Shuiwang Ji. Group equivariant fourier neural operators for partial differential equations. *arXiv preprint arXiv:2306.05697*, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

Peter Holderrieth, Michael J Hutchinson, and Yee Whye Teh. Equivariant learning of stochastic fields: Gaussian processes and steerable conditional neural processes. In *International conference on machine learning*, pages 4297–4307. PMLR, 2021.

Daolang Huang, Manuel Haussmann, Ulpu Remes, ST John, Grégoire Clarté, Kevin Luck, Samuel Kaski, and Luigi Acerbi. Practical equivariances via relational conditional neural processes. *Advances in Neural Information Processing Systems*, 36:29201–29238, 2023.

Saurav Jha, Dong Gong, Xuesong Wang, Richard E Turner, and Lina Yao. The neural process family: Survey, applications and perspectives. *arXiv preprint arXiv:2209.00517*, 2022.

Yohan Jung and Jinkyoo Park. Bayesian convolutional deep sets with task-dependent stationary prior. In *International Conference on Artificial Intelligence and Statistics*, pages 3795–3824. PMLR, 2023.

Makoto Kawano, Wataru Kumagai, Akiyoshi Sannai, Yusuke Iwasawa, and Yutaka Matsuo. Group equivariant conditional neural processes. *arXiv preprint arXiv:2102.08759*, 2021.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Mingyu Kim, Kyeongryeol Go, and Se-Young Yun. Neural processes with stochastic attention: Paying more attention to the context dataset. *arXiv preprint arXiv:2204.05449*, 2022.

David M Knigge, David W. Romero, Albert Gu, Efstratios Gavves, Erik J Bekkers, Jakub Mikolaj Tomczak, Mark Hoogendoorn, and Jan jakob Sonke. Modelling long range dependencies in $n$d: From task-specific to a general purpose CNN. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZW5aK4yCRqU.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021.

Nikola B Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *J. Mach. Learn. Res.*, 24(89):1–97, 2023.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Hyungi Lee, Eunggu Yun, Giung Nam, Edwin Fong, and Juho Lee. Martingale posterior neural processes. *arXiv preprint arXiv:2304.09431*, 2023.

Juho Lee, Yoonho Lee, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, and Yee Whye Teh. Bootstrapping neural processes. *Advances in neural information processing systems*, 33:6606–6615, 2020.

Shibo Li, Xin Yu, Wei Xing, Robert Kirby, Akil Narayan, and Shandian Zhe. Multi-resolution active learning of fourier neural operators. In *International Conference on Artificial Intelligence and Statistics*, pages 2440–2448. PMLR, 2024.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020b.

Ning Liu, Siavash Jafarzadeh, and Yue Yu. Domain agnostic fourier neural operators. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Ning Liu, Yue Yu, Huaiqian You, and Neeraj Tatikola. Ino: Invariant neural operators for learning complex physical systems with momentum conservation. In *International Conference on Artificial Intelligence and Statistics*, pages 6822–6838. PMLR, 2023a.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems*, 2023b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alfred J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3):271–274, 1910. ISSN 0092-7325. doi: 10.1021/j150111a004. URL https://doi.org/10.1021/j150111a004.

Christos Louizos, Xiahan Shi, Klamer Schutte, and Max Welling. The functional neural process. *Advances in Neural Information Processing Systems*, 32, 2019.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

Chao Ma and José Miguel Hernández-Lobato. Functional variational inference based on stochastic process generators. *Advances in Neural Information Processing Systems*, 34:21795–21807, 2021.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In *International Conference on Machine Learning*, pages 4222–4233. PMLR, 2019.

Stratis Markou, James Requeima, Wessel P Bruinsma, Anna Vaughan, and Richard E Turner. Practical conditional neural processes via tractable dependent predictions. *arXiv preprint arXiv:2203.08775*, 2022.

Emile Mathieu, Adam Foster, and Yee Teh. On contrastive representations of stochastic processes. *Advances in Neural Information Processing Systems*, 34:28823–28835, 2021.

Emile Mathieu, Vincent Dutordoir, Michael J Hutchinson, Valentin De Bortoli, Yee Whye Teh, and Richard E Turner. Geometric neural diffusion processes. *arXiv preprint arXiv:2307.05431*, 2023.

Peiman Mohseni, Nick Duffield, Bani Mallick, and Arman Hasanzadeh. Adaptive conditional quantile neural processes. In *Uncertainty in Artificial Intelligence*, pages 1445–1455. PMLR, 2023.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. *arXiv preprint arXiv:2207.04179*, 2022.

Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

Luis A Ortega, Simón Rodríguez Santana, and Daniel Hernández-Lobato. Deep variational implicit processes. *arXiv preprint arXiv:2206.06720*, 2022.

Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters – improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

Md Ashiqur Rahman, Manuel A Florez, Anima Anandkumar, Zachary E Ross, and Kamyar Azizzadenesheli. Generative adversarial neural operators. *arXiv preprint arXiv:2205.03017*, 2022a.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators. *arXiv preprint arXiv:2204.11127*, 2022b.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators. *arXiv preprint arXiv:2204.11127*, 2022c.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019.

Bogdan Raonic, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of pdes. *Advances in Neural Information Processing Systems*, 36: 77187–77200, 2023.

Carl Edward Rasmussen, Christopher KI Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

David W Romero, Anna Kuzina, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Ckconv: Continuous kernel convolution for sequential data. *arXiv preprint arXiv:2102.02611*, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

Tim GJ Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 35: 22686–22698, 2022.

Simón Rodríguez Santana, Bryan Zaldivar, and Daniel Hernández-Lobato. Function-space inference with sparse implicit processes. *arXiv preprint arXiv:2110.07618*, 2021.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn. Sequential neural processes. *Advances in Neural Information Processing Systems*, 32, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. *arXiv preprint arXiv:2111.13802*, 2021.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Anna Vaughan, Will Tebbutt, J Scott Hosking, and Richard E Turner. Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development Discussions*, 2021:1–25, 2021.

Michael Volpp, Fabian Flürenbrock, Lukas Grossberger, Christian Daniel, and Gerhard Neumann. Bayesian context aggregation for neural processes. In *ICLR*, 2021.

V. Volterra. Variazioni e fluttuazioni del bumero d'ondividui in specie animali conviventi. *Memoria della Reale Accademia Nazionale dei Lincei*, 2:31–113, 1926.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the limitations of representing functions on sets. In *International conference on machine learning*, pages 6487–6494. PMLR, 2019.

Martin J Wainwright and Eero Simoncelli. Scale mixtures of gaussians and the statistics of natural images. *Advances in neural information processing systems*, 12, 1999.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European conference on computer vision*, pages 108–126. Springer, 2020.

Qi Wang and Herke Van Hoof. Doubly stochastic variational inference for neural processes with hierarchical latent variables. In *International Conference on Machine Learning*, pages 10018–10028. PMLR, 2020.

Qi Wang and Herke van Hoof. Learning expressive meta-representations with mixture of expert neural processes. *Advances in neural information processing systems*, 35:26242–26255, 2022.

Qi Wang, Marco Federici, and Herke van Hoof. Bridge the inference gaps of neural processes via expectation maximization. In *The Eleventh International Conference on Learning Representations*, 2022.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

Veit David Wild, Robert Hu, and Dino Sejdinovic. Generalized variational inference in function spaces: Gaussian measures meet bayesian deep learning. *Advances in Neural Information Processing Systems*, 35:3716–3730, 2022.

Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam Kosiorek, and Yee Whye Teh. Metafun: Meta-learning with iterative functional updates. In *International Conference on Machine Learning*, pages 10617–10627. PMLR, 2020.

Jin Xu, Emilien Dupont, Kaspar Märtens, Thomas Rainforth, and Yee Whye Teh. Deep stochastic processes via functional markov transition operators. *Advances in Neural Information Processing Systems*, 36:37975–37994, 2023.

Jaesik Yoon, Gautam Singh, and Sungjin Ahn. Robustifying sequential neural processes. In *International Conference on Machine Learning*, pages 10861–10870. PMLR, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.