ThermoPhoton: Fast 3D Thermal Simulation of Photonic Integrated Circuits via Operator Learning

Weilong Guan[1,2†], Li Huang[1†], Yuxuan Lin[1], Yuchao Wu[1], Yeyu Tong[1,2], Yuzhe Ma[1*]
[1]Microelectronics Thrust, HKUST(GZ)
[2]Guangdong-Macao Joint Laboratory for Modular Chip Design and Testing, HKUST(GZ)
{weiguan741,lhuang913,ylin440,ywu092}@connect.hkust-gz.edu.cn {yeyutong,yuzhema}@hkust-gz.edu.cn

2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD) | 979-8-3315-1560-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCAD66269.2025.11240705

***Abstract*—As silicon photonic integrated circuits (PICs) scale in density and integration level, thermal crosstalk significantly impacts chip performance and reliability, necessitating careful thermal-aware design. Traditional numerical solvers are prohibitively slow for large-scale 3D simulation, while existing machine learning surrogates struggle with generalization, especially under the complex distributed heaters and layered structures unique to PICs. We present ThermoPhoton, an operator-learning neural architecture tailored for efficient, accurate 3D thermal modeling of PICs. ThermoPhoton introduces a Pseudo-3D source representation (Pseudo-3D) that leverages device stratification, and applies Zero Coordinate Shift (ZCS) encoding to optimize physics-informed loss computation. Attention mechanisms further enhance the capture of sharp thermal gradients and crosstalk. On industry-standard benchmarks, ThermoPhoton achieves a mean absolute percentage error of 0.07%, reduces peak GPU memory by 67.1%, and shortens training time by 37.9% compared to prior operator-based methods, enabling fast, reliable, and scalable thermal analysis for next-generation photonic chips.**

I. Introduction

Silicon photonic integrated circuits (PICs) are rapidly emerging as a key enabler for high-speed data processing and interconnect, driven by escalating demands in deep learning accelerators, data-center communications, and high-performance computing [1]–[7]. These circuits already demonstrate the capability to integrate diverse photonic functions onto a single chip, providing scalable and cost-effective solutions to meet the growing demands for faster and more efficient data handling. However, as integration scale and density continue to grow, PIC design flows often require substantial expertise and hands-on effort, especially for complex layouts and performance optimization [8], [9]. The management of thermal effects, particularly thermal crosstalk, has become a critical issue that poses a threat to the performance and reliability of photonic circuits due to the presence of thermo-optical effects.

Thermal crosstalk occurs when heat generated by microheaters diffuses to nearby photonic components, causing unintended optical phase shifts that can significantly degrade system performance and must be carefully managed to ensure reliable operation in complex photonic circuits [10]. For instance, in reconfigurable Mach-Zehnder interferometers (MZIs), phase shifts are controlled by microheaters. However, the heat they generate can affect adjacent waveguides, leading to undesired phase changes. Historically, thermal analysis of integrated circuits has been dominated by numerical methods such as the finite-difference method (FDM) and finite-element method (FEM), which solve heat diffusion equations for complex geometries with high accuracy [11]–[14]. These simulations deliver high physical fidelity but are prohibitively computationally intensive, demanding complete re-execution for every design change. Consequently, they are poorly suited for iterative design cycles or real-time optimization. Model-order reduction techniques have accelerated analysis but provide only partial mitigation of this limitation and often rely on hand-crafted simplifications that limit adaptability to new layouts or boundary conditions [15], [16].

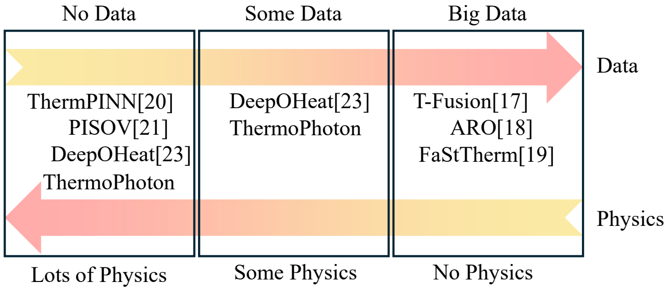

Fig. 1 Categorization of machine learning-based thermal modeling approaches by their use of physical knowledge and data availability.

Machine learning has recently emerged as a compelling alternative for thermal modeling, with the potential to circumvent the high computational cost of traditional numerical simulations. To clarify the landscape of machine learning approaches for this task, we categorize existing methods along two axes: the amount of incorporated physics and the availability of data, as illustrated in Fig. 1. This framework reveals three principal classes: physics-dominated methods, data-driven methods, and hybrid operator learning approaches.

Data-driven approaches, such as T-Fusion [17], ARO [18], and FaStTherm [19], learn complex mappings from structural parameters to thermal behavior and can achieve high accuracy when large datasets are available. However, the need for extensive labeled data makes them less practical for photonics, where each data point is costly to obtain, and these models often lack interpretability. By contrast, physics-informed methods like ThermPINN [20] and PISOV [21] incorporate governing equations directly into the learning process, enabling reasonable accuracy with limited data. Yet, their generalization to complex or unseen layouts is often limited due to their case-specific optimization. To leverage the strengths of both paradigms, hybrid frameworks such as Pi-DeepONet [22] have been developed, integrating operator learning's generalization power with the physics-informed foundations of PINNs. Based on this architecture, Liu *et al.* introduced DeepOHeat [23], which applies Pi-DeepONet to chip-scale thermal simulation and achieves significant acceleration in thermal prediction tasks for VLSI designs.

While operator learning frameworks like DeepOHeat show considerable promise, their direct application to PICs faces challenges due to fundamental differences in thermal modulation. Unlike electronic chips with localized heat sources, thermally tuned PICs feature densely and uniformly distributed microheaters. This distributed heating induces pronounced thermal crosstalk and demands signif-

[†]These authors contributed equally to this work.
[*]Corresponding author

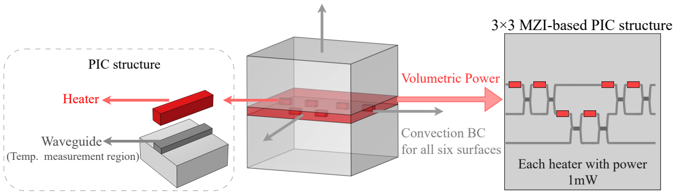

Fig. 2 Schematic of the physical setup for operator learning: the simulation domain is a multi-layer rectangular box with distributed microheaters. Volumetric heat generation is localized within a specific layer, with structure shown for a 3×3 MZI-based photonic neural network. All six surfaces are subject to convection boundary conditions.

icantly higher spatial resolution to resolve temperature gradients accurately. A further challenge arises from the physical structure of PICs. The metal heater and the optical waveguide are fabricated in two closely spaced, yet distinct layers. The primary concern is the thermal impact on the waveguide, which is highly sensitive to temperature fluctuations [24]. Unlike electronic chips, where heat sources frequently coincide with regions of interest and allow for effective two-dimensional modeling, the vertical separation between heaters and waveguides in PICs mandates full three-dimensional thermal simulation to reliably capture interlayer heat transfer.

As illustrated in Fig. 2, the majority of heaters are concentrated within a single plane, while the waveguide of interest resides in a nearby layer. This spatial configuration motivates a critical efficiency-enhancing strategy: we introduce a pseudo-3D source representation (Pseudo-3D), which approximates a 3D heat source by combining a two-dimensional heater map with a depth-wise extension informed by prior knowledge of the PIC structure. By leveraging the physical layer structure of PICs, Pseudo-3D substantially reduces input dimensionality and computational complexity, while preserving the fidelity required for accurate 3D temperature field prediction.

In addition to the above, operator learning frameworks often suffer from low training efficiency, primarily due to the computational cost associated with enforcing physics-informed constraints throughout the spatial domain. To address this, we incorporate the Zero Coordinate Shift (ZCS) technique [25], which eliminates positional redundancy in loss evaluation and enables more efficient enforcement of the governing PDE constraints during training.

Motivated by these considerations, we introduce the following key innovations in the ThermoPhoton framework to address the unique challenges of 3D thermal modeling in PICs:

- We develop ThermoPhoton, a novel operator learning-based neural architecture specifically designed for three-dimensional thermal simulation of photonic integrated circuits. To the best of our knowledge, this is the first operator learning framework tailored to the unique physical and structural requirements of PICs.
- We propose a pseudo-3D source representation that exploits the stratified nature of PIC fabrication, enabling a reduction of the simulation input from three dimensions to two, while preserving essential inter-layer thermal interactions.
- We incorporate the Zero Coordinate Shift encoding scheme, which removes positional redundancy in physics-informed loss evaluation and reduces the computational overhead associated with enforcing PDE constraints.
- ThermoPhoton yields an average reduction of 67.1% in peak GPU memory usage and 37.9% in training time compared to DeepOHeat, while maintaining high accuracy with a mean absolute percentage error of 0.07%.

The rest of the paper is organized as follows: Section II introduces the foundations of thermal modeling and operator learning. Section III describes methods to enabling ThermoPhoton. Section IV presents the experiment results. Section V draws our conclusion.

II. Preliminaries

This section reviews the mathematical foundations of steady-state thermal modeling, as well as the general principles of operator learning for partial differential equations.

Notation: Throughout this paper, we use $\mathbf{u}$ to denote the two-dimensional heater power map (*i.e.*, the lateral distribution $q(x, y)$), and $\mathbf{y} = (x, y, z)$ to denote a spatial query location. The neural operator $\mathcal{G}_\theta$ thus approximates the mapping $\mathbf{u} \mapsto T(\mathbf{y})$ for any $\mathbf{y}$ in the domain.

A. Governing Equation and Boundary Conditions

The temperature distribution $T(x, y, z)$ in a three-dimensional domain $\Omega \subset \mathbb{R}^3$ is governed by the steady-state heat equation,

$$\nabla \cdot (k(x, y, z)\nabla T(x, y, z)) + Q(x, y, z) = 0, \quad (x, y, z) \in \Omega, \tag{1}$$

where $k(x, y, z)$ denotes the spatially dependent thermal conductivity and $Q(x, y, z)$ is the volumetric heat generation rate. The tensorial nature of $k(x, y, z)$ accounts for possible anisotropy in the medium.

The boundary $\partial\Omega$ is subject to a combination of Dirichlet, Neumann, and Robin conditions, which are specified as follows:

$$\begin{aligned} T(x, y, z) &= T_{\text{ext}}(x, y, z), \quad (x, y, z) \in \Gamma_D, \\ -k(x, y, z)\nabla T(x, y, z) \cdot \mathbf{n} &= q_{\text{ext}}(x, y, z), \quad (x, y, z) \in \Gamma_N, \\ -k(x, y, z)\nabla T(x, y, z) \cdot \mathbf{n} &= h(x, y, z)\left(T(x, y, z) - T_\infty(x, y, z)\right), \\ &\quad (x, y, z) \in \Gamma_R. \end{aligned}$$

where $\mathbf{n}$ denotes the outward normal vector, $h(x, y, z)$ is the convection coefficient, and $T_\infty(x, y, z)$ denotes the ambient temperature. The domain boundary is partitioned into disjoint subsets Γ_D, Γ_N, and Γ_R corresponding to Dirichlet, Neumann, and Robin boundaries, respectively.

B. Operator Learning for Partial Differential Equations

Operator learning aims to approximate the solution operator $\mathcal{G}_\theta$ that maps a given input function $\mathbf{u}$ and a query location $\mathbf{y}$ to the solution of the PDE:

$$T(\mathbf{y}) = \mathcal{G}_\theta(\mathbf{u})(\mathbf{y}), \tag{2}$$

where $\mathbf{u}$ represents the input power map, and $\mathbf{y} = (x, y, z)$ denotes the query location.

A prominent neural operator architecture is the Deep Operator Network (DeepONet), which consists of a branch network that encodes the discretized input function $\mathbf{u}$ and a trunk network that encodes the spatial query $\mathbf{y}$. The output is realized as

$$\mathcal{G}_\theta(\mathbf{u})(\mathbf{y}) = \sum_{k=1}^{d} b_k(\mathbf{u})\, t_k(\mathbf{y}), \tag{3}$$

where $b_k(\mathbf{u})$ and $t_k(\mathbf{y})$ are the outputs of branch and trunk networks, respectively, and θ is the set of all trainable parameters.

The model is typically trained by minimizing a loss function that incorporates the residuals of the governing equation and the boundary

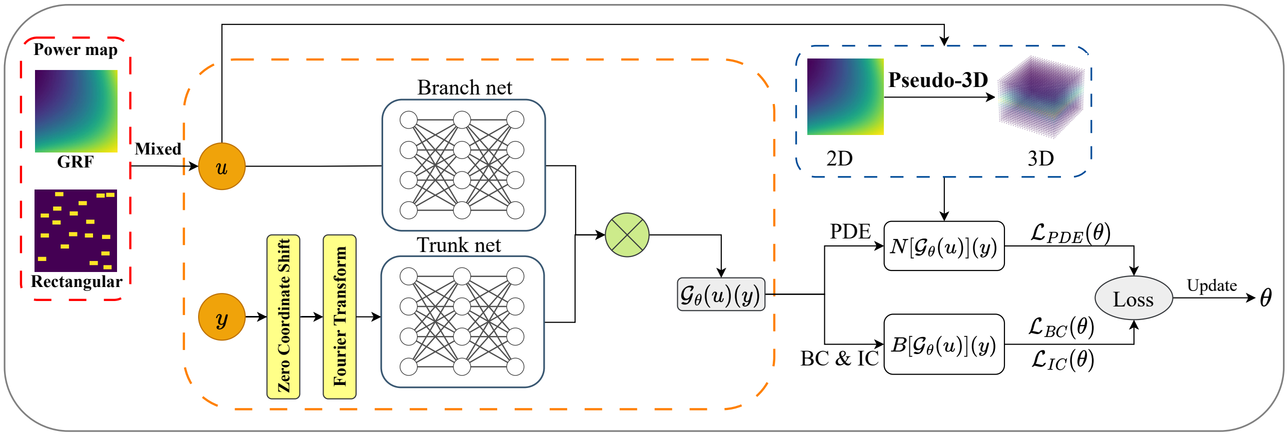
Power map
GRF
Rectangular
Mixed
u
y
Branch net
Trunk net
Zero Coordinate Shift
Fourier Transform
$\mathcal{G}_\theta(u)(y)$
Pseudo-3D
2D
3D
PDE
$N[\mathcal{G}_\theta(u)](y)$
$\mathcal{L}_{PDE}(\theta)$
Loss
Update
θ
BC & IC
$B[\mathcal{G}_\theta(u)](y)$
$\mathcal{L}_{BC}(\theta)$
$\mathcal{L}_{IC}(\theta)$

Fig. 3 Schematic of the ThermoPhoton neural operator architecture. The branch network encodes the input function (the power map $\mathbf{u}$), while the trunk network encodes the query location $\mathbf{y}$. The output is given by a bilinear interaction between the two.

conditions at a set of collocation points. The physics-informed loss for steady-state problems is given by

$$\mathcal{L}_{\text{physics}} = \lambda_{\text{PDE}}\mathcal{L}_{\text{PDE}} + \lambda_{\text{BC}}\mathcal{L}_{\text{BC}}, \tag{4}$$

where $\mathcal{L}_{\text{PDE}}$ quantifies the residual of the governing equation and $\mathcal{L}_{\text{BC}}$ measures the residual of the boundary conditions. The weighting coefficients λ_{PDE} and λ_{BC} control the relative contributions of the two terms.

Automatic differentiation facilitates the evaluation of the required derivatives for loss computation, supporting mesh-independent and data-efficient learning. When available, reference solution data can be incorporated via a supervised loss term, allowing a hybrid training strategy.

The mathematical framework summarized here provides the foundation for the development of specialized operator learning approaches for physical modeling tasks, such as thermal analysis in PICs. For theoretical background and further details, see [22], [26].

III. Thermal Operator Learning Framework

Building upon the theoretical foundations above, we present the ThermoPhoton framework, specifically designed for steady-state thermal modeling in PICs. This section details the problem formulation, neural operator structure, and the physics-informed training strategy.

A. Problem Formulation

The physical scenario considered is illustrated in Fig. 2. The computational domain Ω is modeled as a multi-layer rectangular box, reflecting the typical structure of a PIC. Heat generation is localized within a designated layer containing an array of microheaters. Each heater is assigned a specific position and power, and the lateral heater power distribution is represented as a two-dimensional map $\mathbf{u}(x, y)$, such as the 3×3 MZI-based configuration shown in the Fig. 2.

All boundaries of the simulation domain are imposed with Robin-type (convective) boundary conditions to capture heat exchange with the ambient environment on all six surfaces.

To efficiently represent both the vertical and lateral structure of heat injection, we use a pseudo-3D source representation for the volumetric heat generation:

$$Q(x, y, z) = C\,\mathbf{u}(x, y)\,w(z), \tag{5}$$

where $\mathbf{u}(x, y)$ is the normalized lateral power map, $w(z)$ is the normalized vertical profile, and C is a normalization constant ensuring the prescribed total power.

Given a specified heater pattern $\mathbf{u}$ and device geometry, the objective is to predict the steady-state temperature field $T(x, y, z)$ throughout the domain, subject to these boundary and source conditions.

B. Operator Architecture and Physics-Informed Training

The ThermoPhoton operator architecture, shown schematically in Fig. 3, extends the DeepONet paradigm to chip-scale thermal modeling. The model receives as input the discretized heat source map $\mathbf{u}$ and a spatial query point $\mathbf{y} = (x, y, z)$. The branch network encodes the heater power distribution, while the trunk network processes the spatial query. To enhance spatial representation and gradient computation, positional encoding (such as Fourier features) and the ZCS technique are applied to $\mathbf{y}$ in the trunk network.

The outputs of the branch and trunk networks are combined via a bilinear interaction to produce the predicted temperature:

$$T(\mathbf{y}) = \mathcal{G}_\theta(\mathbf{u})(\mathbf{y}) = \sum_{k=1}^{d} b_k(\mathbf{u})\,t_k(\mathbf{y}), \tag{6}$$

where $b_k(\mathbf{u})$ and $t_k(\mathbf{y})$ are the outputs of the branch and trunk networks, respectively.

To address sharp thermal gradients and strong crosstalk in dense layouts, residual connections and self-attention mechanisms are integrated into both the branch and trunk networks. The Pseudo-3D module reconstructs the full three-dimensional volumetric heat source from the two-dimensional map $\mathbf{u}$, enabling accurate modeling of interlayer effects.

The ThermoPhoton operator is trained in a physics-informed manner by minimizing a loss function that penalizes residuals of the underlying PDE and boundary conditions. Specifically, for each batch, a set of collocation points $\{\mathbf{y}_j\}$ is randomly sampled within the domain and on its boundaries. At each point, the model prediction $T(\mathbf{y}_j) = \mathcal{G}_\theta(\mathbf{u})(\mathbf{y}_j)$ is used to evaluate the residuals of the steady-state heat equation and associated boundary conditions:

$$\mathcal{R}_{\text{PDE}}(\mathbf{y}_j) = \nabla \cdot \big(k(\mathbf{y}_j)\nabla T(\mathbf{y}_j)\big) + Q(\mathbf{y}_j), \tag{7}$$

where $\mathbf{y}_j \in \Omega$.

$$\mathcal{R}_{\text{D}}(\mathbf{y}_j) = T(\mathbf{y}_j) - T_{\text{ext}}(\mathbf{y}_j), \tag{8}$$

where $\mathbf{y}_j \in \Gamma_D$.

$$\mathcal{R}_{\text{N}}(\mathbf{y}_j) = -k(\mathbf{y}_j)\,\nabla T(\mathbf{y}_j) \cdot \mathbf{n} - q_{\text{ext}}(\mathbf{y}_j), \tag{9}$$

where $\mathbf{y}_j \in \Gamma_N$.

$$\mathcal{R}_{\text{R}}(\mathbf{y}_j) = -k(\mathbf{y}_j)\,\nabla T(\mathbf{y}_j)\cdot\mathbf{n} - h(\mathbf{y}_j)\big(T(\mathbf{y}_j) - T_\infty(\mathbf{y}_j)\big), \quad (10)$$

where $\mathbf{y}_j \in \Gamma_R$.

The total physics-informed loss is the weighted mean squared residual over all collocation points:

$$\mathcal{L}_{\text{physics}}(\theta) = \lambda_{\text{PDE}}\,\overline{\mathcal{R}^2_{\text{PDE}}} + \lambda_{\text{D}}\,\overline{\mathcal{R}^2_{\text{D}}} + \lambda_{\text{N}}\,\overline{\mathcal{R}^2_{\text{N}}} + \lambda_{\text{R}}\,\overline{\mathcal{R}^2_{\text{R}}}, \quad (11)$$

where $\overline{\mathcal{R}^2_{\text{PDE}}}$ denotes the mean squared PDE residual over interior points, and similarly for the boundary terms. The weights λ_* are tunable hyperparameters.

Spatial derivatives in the residuals are computed via automatic differentiation, with the ZCS technique ensuring efficient and accurate evaluation even for batched inputs. If reference data is available, a supervised loss term can be incorporated to further enhance accuracy.

During training, the network parameters θ are optimized using a two-stage protocol with the Adam optimizer [27]. In the first stage, a relatively high learning rate is used to enable rapid exploration of the parameter space and efficient minimization of the physics-informed loss. Once the loss plateaus or sufficient progress is achieved, the learning rate is reduced for a second stage of fine-tuning. The strategy successfully mitigates convergence to suboptimal solutions while ensuring robust, high-precision results.

C. Enhanced Network Architecture

To further improve the modeling of complex and non-smooth thermal patterns, ThermoPhoton incorporates advanced deep learning modules. The trunk network and any branch network handling low-dimensional or structured data are implemented as residual multilayer perceptrons (MLPs), where each residual block contains two fully connected layers with nonlinear activation (typically $\tanh$) and a skip connection:

$$\mathbf{h}^{(l+1)} = \tanh\Big(\mathcal{F}_2^{(l)}\big(\tanh(\mathcal{F}_1^{(l)}(\mathbf{h}^{(l)}))\big) + \text{shortcut}(\mathbf{h}^{(l)})\Big), \quad (12)$$

where $\mathcal{F}_1^{(l)}$ and $\mathcal{F}_2^{(l)}$ are fully connected layers, and shortcut is a skip mapping. This design enhances convergence and stability in deep networks [28].

For high-resolution or image-like branch inputs, a convolutional ResNet is used, with each residual block of the form

$$\mathbf{z}^{(l+1)} = \mathbf{z}^{(l)} + \mathcal{G}^{(l)}(\mathbf{z}^{(l)}), \quad (13)$$

where $\mathcal{G}^{(l)}$ denotes a sequence of convolutions and nonlinearities, enabling multiscale spatial feature extraction.

To capture long-range dependencies, a self-attention module [29] follows the ResNet. Feature maps are projected into query, key, and value tensors via 1×1 convolutions, attention weights are computed, and features are aggregated globally, allowing the model to learn both local and global thermal correlations.

D. Pseudo-3D Source Representation

A central challenge in 3D thermal modeling of PICs is the high dimensionality of the internal heat source $Q(x,y,z)$, which leads to significant computational complexity. To address this, we propose a pseudo-3D source representation (Pseudo-3D), which exploits the layered structure of PICs to reduce input dimensionality by expressing the 3D heat source as

$$Q(x,y,z) = C\cdot\mathbf{u}(x,y)\cdot w(z), \quad (14)$$

where $\mathbf{u}(x,y)$ denotes the normalized lateral heatmap, $w(z)$ is the vertical profile, and C is a normalization constant ensuring the total power equals P_{total}:

$$C = \frac{P_{\text{total}}}{\left(\iint_{\Omega_{xy}}\mathbf{u}(x,y)\,dxdy\right)\left(\int_{\Omega_z} w(z)\,dz\right)}. \quad (15)$$

The choice of $w(z)$ is progressively refined to balance physical fidelity with model learnability. The uniform profile,

$$w_{\text{uniform}}(z) = \begin{cases} 1, & z_1 \le z \le z_2 \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

is commonly employed in engineering, as it directly represents scenarios where heat is generated uniformly within a specified thin film or structural layer—a typical outcome of device fabrication.

Nevertheless, the abrupt discontinuities at the layer boundaries can impede convergence and generalization in data-driven models. To alleviate this issue, the Gaussian profile is frequently adopted:

$$w_{\text{Gaussian}}(z) = \exp\left(-\frac{(z-z_0)^2}{2\sigma^2}\right), \quad (17)$$

where z_0 and σ denote the center and width, respectively. The Gaussian function offers a smooth, differentiable approximation of localized heat generation, facilitating numerical optimization and learning.

For further improvement, particularly to capture long-range thermal diffusion, a Lorentzian-like profile can be used:

$$w_{\text{Lorentz}}(z) = \frac{1}{1+\alpha(z-z_0)^2}, \quad (18)$$

where α controls the width and tail decay. This profile provides extended support and smoother transitions, more accurately reflecting the physical characteristics of heat spreading in multilayer structures. This formulation ensures that the total heat input remains consistent regardless of the specific choice of $w(z)$.

E. ZCS-Based Gradient Computation and Implementation

Efficient evaluation of spatial derivatives is critical for enforcing the physics-informed loss in operator learning. We employ the Zero Coordinate Shift method, which improves both memory and computational efficiency when computing gradients via automatic differentiation (AD).

Instead of treating spatial coordinates $\mathbf{y}_j$ as independent leaf variables in AD, we introduce a shared dummy shift variable z and define a perturbed function:

$$v_{ij}(z) := \mathcal{G}_\theta(\mathbf{u}_i, \mathbf{y}_j + z). \quad (19)$$

The gradient with respect to the spatial coordinate is then reparameterized as

$$\left.\frac{\partial\mathcal{G}_\theta}{\partial\mathbf{y}_j}\right|_{\mathbf{y}_j} = \left.\frac{\partial v_{ij}(z)}{\partial z}\right|_{z=0}. \quad (20)$$

To facilitate efficient reverse-mode AD over batched inputs, a scalar-valued root function is constructed:

$$\omega := \sum_{i,j} a_{ij}\cdot v_{ij}(z), \quad (21)$$

where a_{ij} are dummy weights. The final derivative is then obtained via

$$\frac{\partial\mathcal{G}_\theta}{\partial\mathbf{y}_j} = \frac{\partial}{\partial a_{ij}}\left.\left(\frac{\partial\omega}{\partial z}\right)\right|_{z=0}. \quad (22)$$

Algorithm 1 ZCS Gradient Computation via Nested Reverse-Mode AD

```
Require: Input field $\mathbf{u}$, coordinates $\mathbf{y}$, dummy shift $z$, dummy weights $a$
Ensure: Gradients $g_{ij} = \frac{\partial \mathcal{G}_\theta(\mathbf{u},\mathbf{y}_j)}{\partial \mathbf{y}_j}$
Perturb coordinates: $v_{ij}(z) \leftarrow \mathcal{G}_\theta(\mathbf{u}, \mathbf{y}_j + z)$
Construct scalar root: $\omega \leftarrow \sum_{i,j} a_{ij} \cdot v_{ij}(z)$
Inner gradient: $d_z \leftarrow \frac{\partial \omega}{\partial z}\Big|_{z=0}$
Outer gradient: $g_{ij} \leftarrow \frac{\partial d_z}{\partial a_{ij}}$
Return $g_{ij}$
```

This nested differentiation transforms a many-roots–many-leaves gradient problem into a composition of one-root–one-leaf and one-root–many-leaves gradients, both efficiently supported by modern AD frameworks.

The ZCS-based derivative computation is summarized in Algorithm 1. This approach avoids explicit Python loops and preserves a compact computation graph, enabling high efficiency in practice.

F. Training Dataset Design

To enhance generalization in chip-scale thermal modeling, we introduce an inductive bias through hybrid sampling of input fields. Rather than relying solely on smooth Gaussian random fields (GRFs), each training sample $\mathbf{u}^{(i)}(x, y)$ is constructed as a convex combination:

$$\mathbf{u}^{(i)}(x, y) = (1 - \alpha) \cdot \mathbf{u}^{(i)}_{\mathrm{GRF}}(x, y) + \alpha \cdot \mathbf{u}^{(i)}_{\mathrm{rect}}(x, y), \tag{23}$$

where $\alpha \in [0, 1]$ is a mixing coefficient controlling the degree of spatial localization.

Here, the GRF component is sampled from a zero-mean Gaussian process:

$$\mathbf{u}^{(i)}_{\mathrm{GRF}} \sim \mathcal{GP}(0, K((x, y), (x', y'))), \tag{24}$$

with a squared exponential kernel

$$K = \sigma^2 \exp\left(-\frac{(x - x')^2 + (y - y')^2}{2\ell^2}\right). \tag{25}$$

The rectangular component is defined by a sum of indicator functions:

$$\mathbf{u}^{(i)}_{\mathrm{rect}}(x, y) = \sum_{r=1}^{R_i} A_r \cdot \mathbb{1}_{\mathcal{R}_r}(x, y), \tag{26}$$

where each rectangle $\mathcal{R}_r \subset \Omega_{xy}$ is randomly parameterized, and A_r is its amplitude.

Each training instance may also include sampled convection and boundary components, denoted as $h^{(i)}(x, y)$ and $T^{(i)}_D(x, y)$, respectively. The parameter α controls the structural inductive bias: $\alpha = 0$ yields pure GRF (smooth) inputs, $\alpha = 1$ yields fully rectangular (discontinuous) heat maps, and intermediate values produce hybrid structures with both smooth and sharp features. As shown in Fig. 4, the GRF and rectangular heatmaps represent the two extremes of the sampling process, while hybrid samples interpolate between these regimes, enabling the operator to generalize to a wide variety of realistic source distributions.

IV. Experimental Results

In this section, we evaluate our ThermoPhoton on complex PICs that incorporate multiple volumetric heat sources of varied geometric shapes. Such randomly varying heat sources require high-resolution 3D thermal analysis, which poses a significant computational challenge.

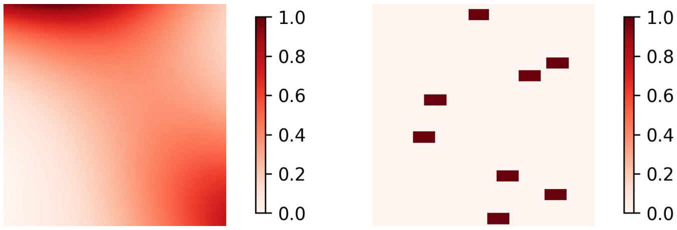

Fig. 4 Examples of sampled heat source maps: (a) a smooth Gaussian random field (GRF, $\alpha = 0$) and (b) a discontinuous rectangular pattern ($\alpha = 1$). Hybrid training samples are generated by convex combination of such fields with mixing coefficient α (see Eq. (23)).

A. Experiment Setup

1) Problem Setup: To develop physically meaningful test cases, we selected PICs composed of MZIs and Microring Resonators (MRRs). In particular, we evaluated two MZI circuits with cuboid heaters ($0.1\,\mathrm{mm} \times 0.05\,\mathrm{mm} \times 0.05\,\mathrm{mm}$) and two MRR circuits with cylindrical heaters (radius $0.05\,\mathrm{mm}$, height $0.05\,\mathrm{mm}$). All circuits were integrated into a PIC, which is modeled as a $120 \times 120 \times 120$ mesh grid-based cubic structure in silicon dioxide (SiO_2) with a uniform isotropic thermal conductivity of 1.4 mW/(mm·K). Convective boundary conditions were applied to all six surfaces of the chip ($HTC = 0.5$ mW/mm^2·K, $T_{\mathrm{amb}} = 298.15$ K).

2) Generating Training Power Maps: We sample all the training powers by a hybrid strategy as mentioned in Section III-F. This approach combines GRF-based distributions with randomly positioned rectangular heat sources to better emulate realistic scenarios. Specifically, 70% of the training samples were derived from 2D GRFs with a length scale of 0.2. The remaining 30% consisted of samples containing 5 to 20 randomly placed rectangular heat sources, each matching the dimensions used in the MZI circuits. A detailed discussion of this strategy and its empirical advantages is provided in Section IV-D2. Corresponding to the 120×120 mesh grid in the heater source layer, each power map was represented as an intensity matrix defined over these coordinates, with dimensions matching the input format of the branch network.

3) ThermoPhoton Settings: Our ThermoPhoton architecture consists of a branch network and a trunk network, both outputting 256-dimensional latent representations. The branch network processed a 120×120 power map and was implemented in two variants. The branch network was based on the standard ResNet-18 architecture, while the Transformer-based branch network extended the same backbone by inserting a self-attention block with 4 heads and a head dimension of 32 after the second stage.

We implemented the trunk network as a residual fully connected (ResFC) module. The spatial coordinates (x, y, z) were first expanded into a 21-dimensional input via a Fourier feature mapping [30] with frequencies 2π, 4π, and 6π. This vector was processed by an input linear layer, then passed through four residual blocks, each containing 128 neurons with $\tanh$ activation, and finally through a projection layer whose output dimension matches that of the branch network.

4) Training Settings: All models were trained for 100,000 iterations using physics-informed loss only, without any supervised temperature data. We employed a two-stage optimization strategy: the first 20,000 iterations use the Adam optimizer with a learning rate of 10^{-3}, followed by 80,000 iterations using AdamW with a learning rate of 10^{-4}. The training was conducted on a single NVIDIA H100 GPU. In each iteration, 30,000 collocation points in the domain and 5,000 boundary points were sampled to evaluate the

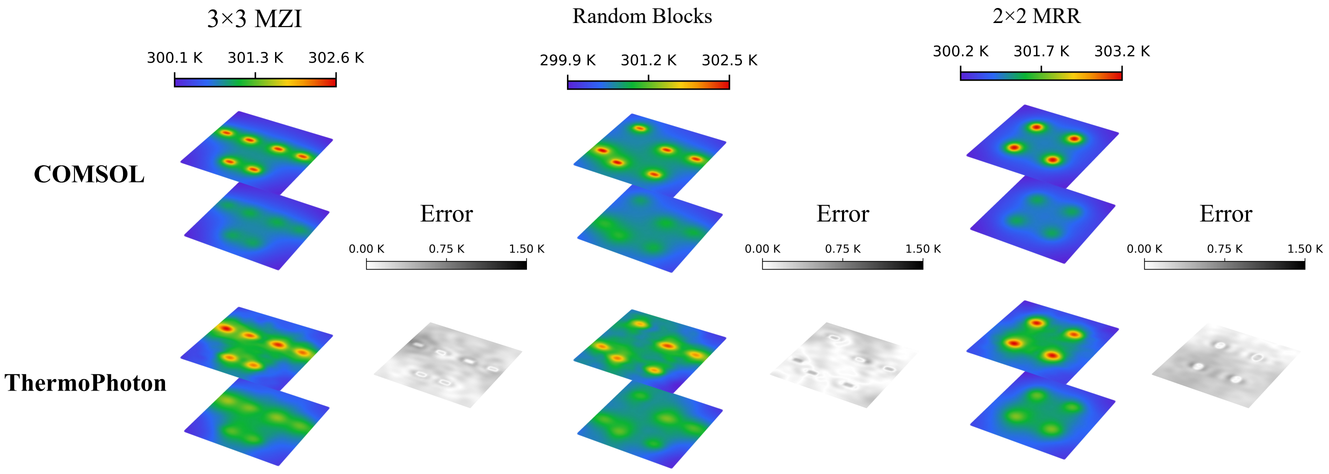

Fig. 5 Transformer-based ThermoPhoton predictions and COMSOL-generated thermal distribution on three representative PIC layouts. Error maps show the absolute error between the prediction and the ground truth.

TABLE I Comparison of ThermoPhoton and DeepOHeat across all test cases in terms of thermal prediction accuracy.

Pattern	**Heaters Num**	**Avg. T_{err} (K)**			**MAPE (%)**			**COMSOL Temp. T (K)**
		Baseline	ResNet	Transformer	Baseline	ResNet	Transformer	
Random Blocks	6	146.084	0.898	0.214	48.37	0.30	0.07	302.014
3 × 3 MZI	6	146.374	0.716	0.109	48.45	0.24	0.04	302.113
4 × 4 MZI	12	147.127	0.778	0.150	48.49	0.26	0.05	303.417
2 × 2 MRR	4	125.013	0.338	0.084	41.30	0.11	0.03	302.694
3 × 3 MRR	9	137.925	1.037	0.445	45.11	0.34	0.15	305.753
Overall	–	140.505	0.753	**0.200**	46.7	0.25	**0.07**	–

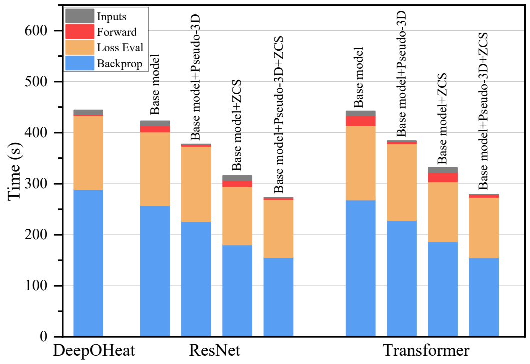

Fig. 6 Training-time breakdown for model variants (per 1000 batches). Each stacked bar shows the total wall-clock time partitioned into four stages: *Inputs* (data loading and transfer to GPU), *Forward* (network forward propagation), *Loss Eval* (PDE loss evaluation), and *Backprop* (gradient computation and parameter update).

PDE and boundary losses. We adopted the ZCS strategy to reduce computational overhead, and dynamically resample collocation points every 4,000 iterations to improve training diversity.

5) Evaluation Metrics: To evaluate the model performance, we compared the predicted temperature fields against reference results generated using COMSOL Multiphysics 6.2. As we focused on PIC thermal analysis, all metrics were computed based on the average temperature in the regions beneath each heater, following the circuit layout specified by the corresponding PDK. We reported the average temperature error (T_{err}) and mean absolute percentage error (MAPE) across all test cases.

TABLE II Peak GPU memory and thermal prediction error.

Method	**GPU Mem. (GB)**	**Avg. T_{err} (K)**	**MAPE (%)**
DeepOHeat	74.1	144.417	47.82
ResNet-based branch network			
Base model	44.1	7.787	2.58
Base model + Pseudo-3D	40.5	0.226	0.08
Base model + ZCS	27.9	2.649	0.88
Base model + Pseudo-3D + ZCS	**24.3**	0.230	0.08
Transformer-based branch network			
Base model	46.0	8.035	2.65
Base model + Pseudo-3D	40.6	0.327	0.11
Base model + ZCS	29.9	0.555	0.18
Base model + Pseudo-3D + ZCS	24.4	**0.212**	**0.07**

B. Comparisons to SOTA Baseline

In recent years, various ML approaches, including ThermPINN, PISOV, and DeepONet-based methods, have been proposed to address complex thermal analysis problems. As discussed in the introduction, direct comparison cannot be made in these studies because of divergent problem definitions in these mainstream approaches. Among them, DeepOHeat aligns most closely with our problem definition, as it employed physics-informed losses directly as the learning objective. Therefore, we adopted DeepOHeat as the baseline to evaluate our proposed ThermoPhoton framework, aiming to highlight the effectiveness of our work. We performed comparative experiments on five PIC-based test cases as shown in TABLE I. For DeepOHeat, we implemented a branch network with 9 fully connected layers (256 neurons per layer) and a trunk network consisting of 9 fully

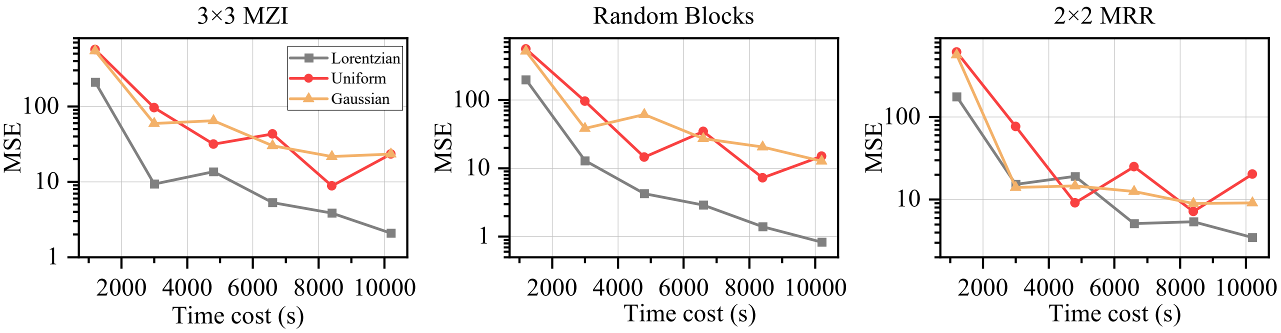

Fig. 7 Comparison of training performance under Lorentzian-based, Gaussian-based, and Uniform-based Pseudo-3D source representation across representative PIC cases.

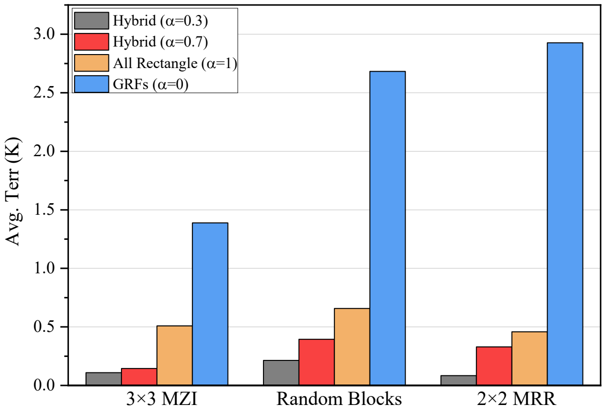

Fig. 8 Average temperature prediction error under different sampling strategies.

connected layers (128 neurons per layer). As shown in TABLE I, DeepOHeat yielded poor accuracy across all test cases, with an average MAPE approaching 47% due to its inadequate handling of 3D heat sources. In contrast, our proposed ThermoPhoton framework achieved significantly better performance, particularly when equipped with Transformer-based branch networks. Fig. 5 further highlights this advantage through visual comparisons, where the Transformer-based model achieved an overall MAPE of only 0.07% and an average error of 0.20 K.

C. Overhead Evaluation

To evaluate the computational efficiency and scalability of ThermoPhoton, we compared four model variants based on ResNet and Transformer branch networks. The runtime breakdown is illustrated in Fig. 6, and the peak memory usage and prediction accuracy are reported in TABLE II. All models were trained for 70,000 iterations with the same condition on a NVIDIA H100 GPU and tested in the 3×3 MZI circuit. Compared to DeepOHeat, which uses MLPs, our base models adopt CNN-based branches, substantially reducing parameter count and GPU memory usage even before applying further optimization. We then applied the Pseudo-3D and ZCS-based gradient computation strategies to improve accuracy, lower peak memory usage, and decrease training runtime. When combined, Pseudo-3D and ZCS delivered the most efficient and accurate configuration, yielding an average 67.1% reduction in peak GPU memory usage and 37.9% reduction in training time across both architectures, relative to DeepOHeat.

D. Hyperparameter Sensitivity Analysis

1) Validation of Pseudo-3D Source Representation: We evaluated our Pseudo-3D by comparing three distributions, Lorentzian, Gaussian, and Uniform, for projecting 2D power maps into 3D. Each model was trained for 20,000 iterations, and models were recorded at fixed time intervals. MSE was used across three representative test cases to evaluate performance. As shown in Fig. 7, the Lorentzian distribution led to faster convergence due to its algebraic decay ($1/(1+z^2)$), thereby better preserving the mid-to-long-range frequency components required for heat transfer modeling. Ultimately, the Lorentzian-based Pseudo-3D improved prediction accuracy by approximately 50% and 80% compared to the Gaussian-based and Uniform-based Pseudo-3D, respectively.

2) Validation of Hybrid Sampling Strategy for Training Heat Source: To validate the advantage of the hybrid sampling training heat source dataset strategy, we conducted comparative experiments under different hybrid sampling ratios and assessed prediction accuracy in three cases. As shown in Fig. 8, hybrid datasets ($0<\alpha<1$) significantly reduced average temperature errors compared to GRFs-only ($\alpha=0$) or rectangle-only ($\alpha=1$) training. This analysis suggests that a hybrid sampling strategy ($\alpha=0.3$) can enhance model generalization, resulting in an average error reduction of 0.406 K and 2.197 K across the three test cases, compared to rectangle-only and GRFs-only training, respectively.

V. Conclusions

We propose ThermoPhoton, a novel physics-informed DeepONet framework for high-resolution 3D thermal modeling of PICs. By integrating two key techniques, Pseudo-3D Source Representation and Zero Coordinate Shift. These enhancements led to a 67.1% reduction in peak GPU memory usage and a 37.9% reduction in training time compared to baseline DeepOHeat, while maintaining high accuracy with a mean absolute percentage error of just 0.07%. In future work, we aim to expand its applicability to broader PIC design tasks, including thermal-aware placement, and plan to explore more advanced training strategies to further enhance model accuracy and reduce training overhead, particularly in large-scale or time-sensitive deployment scenarios.

Acknowledgment

This work is supported in part by the Nansha District Key Area S&T Scheme (No. 2024ZD007) and the Guangdong Science and Technology Department (No.2025B1212150003).

References

[1] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature photonics*, 2017.

[2] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," *Nature Photonics*, 2021.

[3] K. Lu, Z. Chen, H. Chen, W. Zhou, Z. Zhang, H. K. Tsang, and Y. Tong, "Empowering high-dimensional optical fiber communications with integrated photonic processors," *Nature Communications*, 2024.

[4] A. Eldebiky, B. Li, and G. L. Zhang, "Nearuni: Near-unitary training for efficient optical neural networks," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023.

[5] R. Qiu, A. Eldebiky, G. Li Zhang, X. Yin, C. Zhuo, U. Schlichtmann, and B. Li, "Oplixnet: Towards area-efficient optical split-complex networks with real-to-complex data assignment and knowledge distillation," in *Proc. DATE*, 2024.

[6] S. Hua, E. Divita, S. Yu, B. Peng, C. Roques-Carmes, Z. Su, Z. Chen, Y. Bai, J. Zou, Y. Zhu *et al.*, "An integrated large-scale photonic accelerator with ultralow latency," *Nature*, 2025.

[7] S. R. Ahmed, R. Baghdadi, M. Bernadskiy, N. Bowman, R. Braid, J. Carr, C. Chen, P. Ciccarella, M. Cole, J. Cooke *et al.*, "Universal photonic artificial intelligence acceleration," *Nature*, 2025.

[8] W. Bogaerts and L. Chrostowski, "Silicon photonics circuit design: methods, tools and challenges," *Laser & Photonics Reviews*, 2018.

[9] Y. Wu, X. Yu, H. Chen, Y. Luo, Y. Tong, and Y. Ma, "PICBench: Benchmarking LLMs for photonic integrated circuits design," *2025 Design, Automation & Test in Europe Conference (DATE)*, 2025.

[10] M. Orlandin, A. Cem, V. Curri, A. Carena, F. Da Ros, and P. Bardella, "Thermal crosstalk effects in a silicon photonics neuromorphic network," in *2023 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD)*, 2023.

[11] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra, "Efficient full-chip thermal modeling and analysis," in *IEEE/ACM International Conference on Computer Aided Design, 2004. ICCAD-2004.*, 2004.

[12] Z. Liu, S. Swarup, S. X.-D. Tan, H.-B. Chen, and H. Wang, "Compact lateral thermal resistance model of tsvs for fast finite-difference based thermal analysis of 3-d stacked ics," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2014.

[13] Y.-C. Chen, S. Ladenheim, H. Kalargaris, M. Mihajlović, and V. F. Pavlidis, "Computationally efficient standard-cell fem-based thermal analysis," in *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2017.

[14] H. Sultan, A. Chauhan, and S. R. Sarangi, "A survey of chip-level thermal simulators," *ACM Computing Surveys (CSUR)*, 2019.

[15] T.-Y. Wang and C. C.-P. Chen, "Spice-compatible thermal simulation with lumped circuit modeling for thermal reliability analysis based on modeling order reduction," in *International Symposium on Signals, Circuits and Systems. Proceedings, SCS 2003.(Cat. No. 03EX720)*, 2004.

[16] J. Xie and M. Swaminathan, "System-level thermal modeling using nonconformal domain decomposition and model-order reduction," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 2013.

[17] B. Zhang, W. Xing, X. Zhao, and Y. Sun, "T-fusion: Thermal modeling of 3d ics with multi-fidelity fusion," in *Proceedings of the 30th Asia and South Pacific Design Automation Conference*, 2025.

[18] M. Wang, Y. Cheng, W. Zeng, Z. Lu, V. F. Pavlidis, and W. Xing, "Aro: Autoregressive operator learning for transferable and multi-fidelity 3d-ic thermal analysis with active learning," in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024.

[19] T. Zhu, Q. Wang, Y. Lin, R. Wang, and R. Huang, "Fasttherm: Fast and stable full-chip transient thermal predictor considering nonlinear effects," in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024.

[20] L. Chen, J. Lu, W. Jin, and S. X.-D. Tan, "Fast full-chip parametric thermal analysis based on enhanced physics enforced neural networks," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023.

[21] L. Chen, W. Zhu, M. Tang, S. X.-D. Tan, J.-F. Mao, and J. Zhang, "Pisov: Physics-informed separation of variables solvers for full-chip thermal analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[22] S. Wang, H. Wang, and P. Perdikaris, "Learning the solution operator of parametric partial differential equations with physics-informed deeponets," *Science advances*, 2021.

[23] Z. Liu, Y. Li, J. Hu, X. Yu, S. Shiau, X. Ai, Z. Zeng, and Z. Zhang, "Deepoheat: operator learning-based ultra-fast thermal simulation in 3d-ic design," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*, 2023.

[24] I. Teofilovic, A. Cem, D. Sanchez-Jacome, D. Pérez-López, and F. Da Ros, "Thermal crosstalk modelling and compensation methods for programmable photonic integrated circuits," *Journal of Lightwave Technology*, 2024.

[25] K. Leng, M. Shankar, and J. Thiyagalingam, "Zero coordinate shift: Whetted automatic differentiation for physics-informed operator learning," *Journal of Computational Physics*, 2024.

[26] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, "Learning nonlinear operators via deeponet based on the universal approximation theorem of operators," *Nature machine intelligence*, 2021.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[30] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in neural information processing systems*, 2020.