LIGHT-AI INTERACTION: BRIDGING PHOTONICS AND AI WITH CROSS-LAYER HARDWARE/ALGORITHM CO-DESIGN

Jiaqi Gu [1 2] **Hanqing Zhu** [1] **Chenghao Feng** [1 3] **Ray T. Chen** [1] **David Z. Pan** [1]

ABSTRACT

In the post Moore's era, conventional electronic digital computers have encountered escalating challenges in supporting massively parallel and energy-hungry artificial intelligence (AI) workloads, which raises a high demand for a revolutionary computing solution. Optical neural network (ONN) is a promising hardware platform that could represent a paradigm shift with its ultra-fast speed, high parallelism, and low energy consumption. However, designing photonic computing hardware encounters significant challenges in area scalability, noise robustness, adaptability, and design efficiency. In this paper, we present a holistic solution with state-of-the-art cross-layer co-design methodologies towards scalable, robust, and self-learnable integrated photonic neural accelerator designs across the circuit, architecture, and algorithm levels. We will introduce novel ONN architectures with customized circuit and device designs to enable NN inference acceleration, efficient ONN on-chip training algorithms that enable self-learnable photonic AI engines, and AI-assisted automated photonic integrated circuit (PIC) design methodology to form a virtuous cycle of photonics for AI and AI for photonics. Our proposed photonic AI design stack is integrated into our open-source PyTorch-centric ONN library TorchONN to construct customized photonic AI engine designs with high-performance training and optimization facilities.

1 INTRODUCTION

Conventional computing solutions of digital electronics have become a limiting factor in certain domains, most notably intelligent information processing. The proliferation of big data and artificial intelligence (AI) has motivated the investigation of *next-generation specialized AI hardware* to support low-power, low-latency machine intelligence. In recent years, AI computing platforms based on *integrated neuromorphic photonics* are booming due to the ultra-high bandwidth, sub-nanosecond latency, and sub-fJ/MAC energy efficiency of optics. The current optical-electronic hybrid computing hardware can already realize state-of-the-art (SoTA) energy efficiency of around 10 TOPS/W. The theoretical limits of full-optical chips can further boost the efficiency to 1 million TOPS/W, which is 5 orders of magnitude higher than the current SoTA platforms. Such emerging analog photonic AI hardware can make transformative impacts in future datacenters, automotive, military applications, smart sensing, and intelligent edge, enabling foundational breakthroughs in real-time perception, control, decision-making, and learning. The early research efforts focus on diffractive free-space optical computing, optical reservoir computing (Brunner et al., 2013), and spike processing (Tait et al., 2014; Rosenbluth et al., 2009) to achieve optical multi-layer perceptrons (MLPs). Recently, the integrated optical neural networks (ONNs) have attracted extensive research interest given their compactness, energy efficiency, and electronics compatibility (Shen et al., 2017; Cheng et al., 2020; Wetzstein et al., 2020; Shastri et al., 2021), including coherent photonic tensor cores based on broadband devices, e.g., MZIs (Shen et al., 2017), phase shifters (Feng et al., 2022), MMIs, star-couplers (Zhu et al., 2022), and metalens (Wang et al., 2022), and multi-wavelength designs based on micro-ring resonators (Tait et al., 2017; Liu et al., 2019; Gu et al., 2021b), frequency microcomb (Xu et al., 2021), and phase change materials (Miscuglio & Sorger, 2020; Feldmann et al., 2021).

However, designing such emerging photonic AI accelerators have several major challenges that may hinder the practical application of optical AI in the real world. The first challenge is the area scalability due to the large spatial footprint of photonic devices. The packing density of current photonic integrated circuits (PICs) is not comparable to electronic digital chips. More advances in customized devices and novel circuit design methodologies are needed to achieve breakthroughs in scalability. The second chal-

[1]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA. [2]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA. [3]Alpine Optoelectronics, Fremont, CA, USA.. Correspondence to: Jiaqi Gu <jiaqi.gu@asu.edu>.

Accepted to the MLSys Conference SNAP Workshop, Miami, FL, USA, 2023. Copyright 2023 by the author(s).

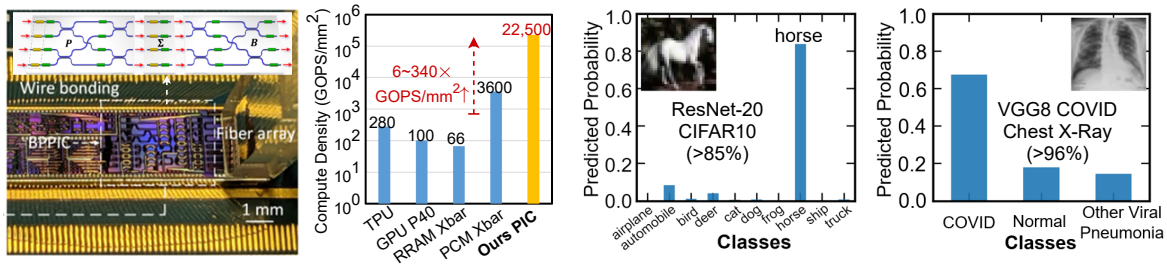

Figure 1: *Left 1*: Our butterfly-style photonic neural chip (Gu et al., 2020b;c; Feng et al., 2022). *Left 2*: Compute density comparison. *Right 1 and 2*: Measured predictions on CIFAR-10 and Chest X-ray Covid (Chowdhury et al., 2020) detection tasks.

lenge is the reliability concerns of photonic analog computing. Due to process variation, environmental changes, and various on-chip noises, the fidelity of the computing results from the photonic AI engines is rather limited. Device-circuit-model co-optimization is necessary to minimize its sensitivity to noises and variations toward robust and reliable photonic AI hardware. Another critical challenge for current optical AI is the training difficulty in adapting the computing engine to changing workloads and environments. The lack of local learnability or self-learnability could disable many important on-device learning applications for the future intelligent edge, e.g., lifelong learning, transfer learning, online adaptation, etc. Moreover, the design efficiency of photonic computing hardware is often limited by traditional manual design flows, which lack intelligence or automation to speed up the development closure and explore the huge design space for better design quality. The key solution to resolving the above challenges is cross-layer hardware/algorithm co-design and intelligent design automation. The following section will provide an overview of light-AI interaction and a full-stack automated co-design solution, including (1) specialized photonic AI hardware designs, (2) scalable on-chip training frameworks, and (3) applying ML for future photonic hardware design automation flow.

2 Scalable, Robust, and Adaptive Photonic AI Platform

2.1 Hardware-Efficient Butterfly-Style Photonic Neural Accelerator

Integrated photonic processors (Shen et al., 2017) have been demonstrated to accelerate general matrix multiplication (GEMM), targeting a photonic substitution of GPUs/TPUs. However, the large spatial footprint of photonic circuits is the bottleneck for further scaling. Besides the continuous miniaturization from device shrinking, we propose to push the limit of scalability by designing specialized photonic circuits to trade redundant matrix expressivity for higher hardware efficiency. To avoid using quadratically many MZIs to build a universal programmable linear unit for GEMM, we break the large MZI structure into basic components, i.e., couplers and phase shifters, and construct a compact photonic neural engine with a butterfly-style circuit topology that significantly **cuts down the optical device usage and realizes similar functionality** (Gu et al., 2020b;c; Feng et al., 2022). Our programmable butterfly-style photonic circuits can realize circuit matrix multiply, Hadamard matrix multiply, and over 64% arbitrary matrices to cover a large enough matrix space for high-accuracy machine learning tasks. We taped out a programmable electronic-photonic neural chip at Advanced Micro Foundry, shown in Fig. 1. With specially designed hardware-aware training methodologies, our chip can implement ResNet-20 and reliably achieve >85% accuracy on the CIFAR-10 image recognition dataset requiring only 3-bit voltage control precision on the diagonal matrix. We further evaluate our photonic chips on a Covid-19 detection task based on chest X-ray images and realize over 96% accuracy. A single 4×4 photonic tensor core can achieve 225 TOPS/mm^2 compute density and 9.5 TOPS/W energy efficiency, which is orders-of-magnitude more powerful than modern GPUs/TPUs, 2-4× more compact and 5-13× shorter in optical delay than the SoTA MZI-based photonic tensor cores.

2.2 Ultra-Compact Photonic Neurons with Customized Multi-Operand Devices

The compute density and energy efficiency of conventional ML accelerators is typically upper-bounded by 1 multiply-accumulate operation (MAC) per device. Moreover, the system performance is often limited by the separate nonlinear activation circuity. To break through this long-lasting performance bottleneck, we propose to **fuse tensor operations and nonlinearity in a single device**. For the first time, we squeeze an 8×8 matrix multiplication into a single 10×10 μm^2 multi-operand microring resonator (MORR) (Gu et al., 2021b; 2022a; Feng et al., 2023). We fully leverage the physics principle of the microring resonators and apply multiple independent control signals to the ring structure to achieve a vector-vector dot-product. Besides, the transmission of the device naturally supports

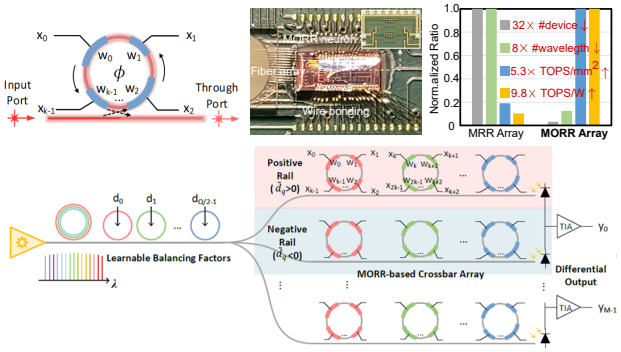

Figure 2: Our MORR-based photonic tensor core schematic and tape-out (Gu et al., 2021b; 2022a). Compared to the standard MRR weight bank, our MORR array achieves significant improvement in device usage, wavelength usage, compute density, and energy efficiency.

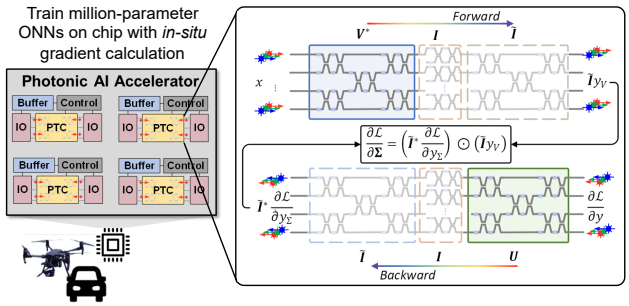

Figure 3: Our scalable ONN on-chip training framework with *in-situ* gradient calculation (Gu et al., 2021d).

built-in reconfigurable nonlinearity. Such nonlinearity can also be learned during training and dynamically reconfigured by tuning the device transmission, which significantly boosts the versatility and expressivity of the designed neurons. Our design can be scaled up by using efficient optimization strategies with structurally sparse matrices and hardware-aware training recipes. Compared to previous photonic tensor cores based on standard microring (MRR) arrays (Tait et al., 2017; Liu et al., 2019), we can realize comparable ML task performance with **quadratically fewer devices**, 8× fewer wavelengths, 5.3× higher compute density, 9.8× higher energy efficiency, and a 63.5% reduction in the simulated system energy consumption. Our team **taped out this MORR-based photonic neuron using AIM Photonics foundry**, shown in Fig. 2. This new design methodology implies an exciting research direction of *neuromorphic computing using customized multi-operand devices*, which shows great potential to push the compute density and efficiency to the extreme.

2.3 Self-Learnable Photonic Neural Accelerator with Efficient On-Chip Training

Besides inference acceleration, future AI systems, especially the intelligent edge, require on-device self-learnability. A self-learnable photonic computing system can (1) address the robustness issues *in situ* and closes the performance gap between simulation and physical deployment; (2) help with data privacy with local learning capability; (3) allow online learning and real-time adaptation on the edge with reduced communication cost; and (4) significantly reduce training energy consumption. Previously, training of photonic neural networks is often offloaded to digital computers with rigorous noise simulation, which is not efficient and usually suffers performance degradation after deployment due to the simulation-reality gap. Generic gradient-free optimization methods, e.g., evolutionary algorithms and brute-force device tuning (Zhou et al., 2020; Shen et al., 2017), have been applied to optimize the photonic circuit parameters, which show limited scalability and stability to handle large-scale ONN training tasks. The adjoint variable method (Hughes et al., 2018a) was introduced to calculate the gradients *in-situ* using per-device optical field monitors. We propose a series of efficient ONN on-chip training protocols to break through the training scalability and efficiency, assuming the input/output observability without access to intermediate circuit states. We propose a series of on-chip training protocols FLOPS, (Gu et al., 2020a), MixedTrain (Gu et al., 2021a), and L^{2}ight (Gu et al., 2021d) to enable self-learnable photonic AI chips with unprecedented training efficiency. Our FLOPS framework is a forward-only zeroth-order optimization flow to enable on-chip training of 1,000 MZIs with a built-in crosstalk handling mechanism, which shows 10× higher training scalability and 4× faster training speed than previous evolutionary and brute-force device tuning methods. We further enhance the solution to a mixed-training framework MixedTrain (Gu et al., 2021a). We partition the photonic circuits into passive and active regions and only train a small subset of active devices in each iteration to reduce parameter update costs. A power optimization technique that prefers low-power parameters is embedded in our sparse zeroth-order coordinate descent optimizer to reduce power consumption. We can further boost the training scalability by 2.5× with over 90% training energy cost reduction. As shown in Fig. 3, to enable million-parameter ONN on-chip training, we propose a subspace optimization algorithm and a multi-level sparse training method L^{2}ight (Gu et al., 2022b) to enable *in-situ* first-order partial gradient calculation and thus enable on-chip training of million-parameter ONNs with 1,000× scalability breakthrough and 30× training cost reduction, enabling efficient self-calibration, online task transfer, lifelong learning, and edge training applications.

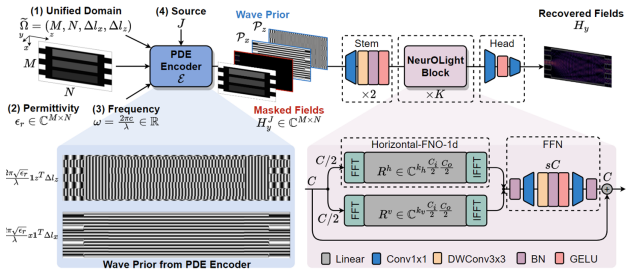

Figure 4: Neural operator-based Maxwell equation solving framework for ultra-fast 2-D optical device simulation.

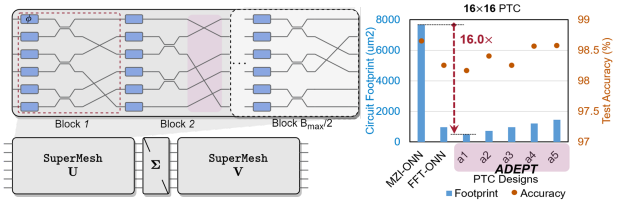

Figure 5: Our automated PIC topology search framework finds more compact designs than prior manual designs.

3 AI-Assisted Intelligent Photonic Hardware Design Automation

3.1 AI-Assisted Photonic Device Simulation

AI-assisted photonic device simulation is a critical step to closing the loop of *light-AI interaction*. Besides using standard devices that already have a compact transfer matrix, optical AI shows a trend to exploit customized photonic structures for scalable optical computing (Gu et al., 2021c; Sunny et al., 2021; Zhu et al., 2022; Wang et al., 2022). Customized devices usually do not have analytical transfer functions. Understanding their behavior heavily relies on numerical simulators (Hughes et al., 2018b) to solve Maxwell partial differential equations (PDEs) to obtain the optical field distribution. The time-consuming optical simulation makes it intractable to perform large-scale outer-loop optimization. Most prior work still uses conventional NNs to predict several key properties based on a few design variables (Tahersima et al., 2019; Trivedi et al., 2019), which is an ad-hoc function approximator without learning the light propagation property. Several works attempt to leverage physics-informed NNs (PINNs) (Tang et al., 2022; Chen et al., 2021; Lim & Psaltis, 2022) to predict electromagnetic field solutions, which requires nontrivial implementation efforts on Maxwell equations and boundary conditions. To learn *a family of parametric Maxwell PDEs* that models the *joint probability* of different PDE variables, we propose a physics-agnostic light field prediction framework `NeurOLight`, shown in Fig. 4. We propose a joint PDE encoder with wave prior and masked source modeling for compact PDE representation. Our lightweight cross-shaped `NeurOLight` backbone design achieves a superior balance between modeling capability and parameter efficiency. In addition, our superposition-based mixup technique significantly boosts the data efficiency and model generalizability. Experiments show that `NeurOLight` outperforms prior DNN models with 53.8% better prediction fidelity and 44.2% less parameter cost, serving as an over 200× faster surrogate model to the numerical solvers in photonic device simulation.

3.2 Automated Photonic Circuit Design Flow

Previous photonic tensor cores (PTCs) are all hand-designed based on matrix decomposition theory (Shen et al., 2017; Feng et al., 2022), which leaves a large design space unexplored with unsatisfying design quality and lacks the adaptability to meet various device specifications and hardware constraints. We observe strong demand for an automatic, efficient, and flexible PTC design methodology. However, PTC design search encounters unique and difficult challenges. The PTC circuit topology has an extremely large and highly discrete search space, which casts *significant optimization difficulties* that prevent the direct application of any combinatorial optimization methods or off-the-shelf neural architecture search methods.

To handle those challenges, we propose the *first* automatic differentiable search framework for photonic tensor core topology design, shown in Fig. 5. Our target is, given certain footprint constraints, we can efficiently search for a photonic circuit topology with good matrix *representability*, compact *footprint*, and high noise *robustness*. Our `ADEPT` constructs a probabilistic photonic `SuperMesh`, employs an augmented Lagrangian method to learn waveguide connections, and adopts binarization-aware training to search coupler locations. With a probabilistic footprint penalty method, `ADEPT` integrates circuit area constraints into `SuperMesh` training procedure to adapt the PTC to various device specifications and footprint constraints. Extensive experiments show the superior flexibility of `ADEPT` for automated PTC topology search adaptive to foundry PDKs. The searched PTC design outperforms prior manual designs with competitive expressiveness, **2×-30×** smaller footprint, and superior robustness. `ADEPT` opens a new paradigm in photonic neurocomputing by "nurturing" photonic circuit design via AI and automation.

4 Conclusion

We present a holistic co-design solution to address the key challenges in the photonic AI design stack and form a virtuous cycle of photonics for AI and AI for photonic hardware design automation. Through hardware customization, we can significantly reduce the area cost and power con-

sumption and boost the noise tolerance of photonic neural networks. To enable self-learnable and adaptive photonic AI accelerators, we introduce a series of scalable on-chip training protocols that enable efficient learning on the edge and adapt to changing and non-ideal environments. To further maximize the design productivity and quality, we presented an AI-assisted photonic device simulation framework and optimization-based automated photonic circuit topology search flow to achieve beyond human design efficiency and hardware performance.

REFERENCES

Brunner, D., Soriano, M. C., Mirasso, C. R., et al. Parallel photonic in- formation processing at gigabyte per second data rates using transient states. *Nature Communications*, 2013.

Chen, M., Lupoiu, R., Mao, C., Huang, D.-H., Jiang, J., Lalanne, P., and Fan, J. Physics-augmented deep learning for high-speed electromagnetic simulation and optimization. *Nature*, 2021.

Cheng, Q., Kwon, J., Glick, M., Bahadori, M., Carloni, L. P., and Bergman, K. Silicon Photonics Codesign for Deep Learning. *Proceedings of the IEEE*, 2020.

Chowdhury, M., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M., Mahbub, Z., Islam, K., Khan, M., Iqbal, A., Al-Emadi, N., Reaz, M., and Islam, M. T. Can AI help in screening Viral and COVID-19 pneumonia? *IEEE ACCESS*, 2020.

Feldmann, J., Youngblood, N., Karpov, M., Gehring, H., Li, X., Stappers, M., Gallo, M. L., Fu, X., Lukashchuk, A., Raja, A., Liu, J., Wright, D., Sebastian, A., Kippenberg, T., Pernice, W., and Bhaskaran, H. Parallel convolutional processing using an integrated photonic tensor core. *Nature*, 2021.

Feng, C., Gu, J., Zhu, H., Ying, Z., Zhao, Z., Pan, D. Z., and Chen, R. T. A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning. *ACS Photonics*, 2022.

Feng, C., Tang, R., Gu, J., Zhu, H., Pan, D. Z., and Chen, R. T. Optically-Interconnected, Hardware-Efficient, Electronic-Photonic Neural Network using Compact Multi-Operand Photonic Devices. In *SPIE Photonics West*, January 2023.

Gu, J., Zhao, Z., Feng, C., Li, W., Chen, R. T., and Pan, D. Z. FLOPS: Efficient On-Chip Learning for Optical Neural Networks Through Stochastic Zeroth-Order Optimization. In *Proc. DAC*, 2020a.

Gu, J., Zhao, Z., Feng, C., et al. Towards area-efficient optical neural networks: an FFT-based architecture. In *Proc. ASPDAC*, 2020b.

Gu, J., Zhao, Z., Feng, C., et al. Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability. *IEEE TCAD*, 2020c.

Gu, J., Feng, C., Zhao, Z., Ying, Z., Chen, R. T., and Pan, D. Z. Efficient on-chip learning for optical neural networks through power-aware sparse zeroth-order optimization. In *Proc. AAAI*, 2021a.

Gu, J., Feng, C., Zhao, Z., Ying, Z., Liu, M., Chen, R. T., and Pan, D. Z. SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators. In *Proc. DATE*, February 2021b.

Gu, J., Zhao, Z., Feng, C., Ying, Z., Chen, R. T., and Pan, D. Z. O2NN: Optical Neural Networks with Differential Detection-Enabled Optical Operands. In *Proc. DATE*, February 2021c.

Gu, J., Zhu, H., Feng, C., Jiang, Z., Chen, R. T., and Pan, D. Z. L2ight: Enabling On-Chip Learning for Optical Neural Networks via Efficient in-situ Subspace Optimization. In *Proc. NeurIPS*, 2021d.

Gu, J., Feng, C., Zhu, H., Zhao, Z., Ying, Z., Liu, M., Chen, R. T., and Pan, D. Z. queezeLight: A Multi-Operand Ring-Based Optical Neural Network with Cross-Layer Scalability. *IEEE TCAD*, July 2022a.

Gu, J., Gao, Z., Feng, C., Zhu, H., Chen, R. T., Boning, D. S., and Pan, D. Z. NeurOLight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation. In *Proc. NeurIPS*, 2022b.

Hughes, T. W., Minkov, M., Shi, Y., and Fan, S. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica*, 2018a.

Hughes, T. W., Minkov, M., Williamson, I. A. D., and Fan, S. Adjoint method and inverse design for nonlinear nanophotonic devices. *ACS Photonics*, 2018b.

Lim, J. and Psaltis, D. Maxwellnet: Physics-driven deep neural network training based on maxwell's equations. *Appl. Phys. Lett.*, 2022.

Liu, W., Liu, W., Ye, Y., Lou, Q., Xie, Y., and Jiang, L. Holylight: A nanophotonic accelerator for deep learning in data centers. In *Proc. DATE*, 2019.

Miscuglio, M. and Sorger, V. J. Photonic tensor cores for machine learning. *Applied Physics Review*, 2020.

Rosenbluth, D., Kravtsov, K., Fok, M. P., et al. A high performance photonic pulse processing device. *Opt. Express*, 17(25), Dec 2009.

Shastri, B. J., Tait, A. N., de Lima, T. F., Pernice, W. H. P., Bhaskaran, H., Wright, C. D., and Prucnal, P. R. Photonics for Artificial Intelligence and Neuromorphic Computing. *Nature Photonics*, 2021.

Shen, Y., Harris, N. C., Skirlo, S., et al. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 2017.

Sunny, F., Mirza, A., Nikdast, M., and Pasricha, S. Crosslight: A cross-layer optimized silicon photonic neural network accelerator. In *Proc. DAC*, 2021.

Tahersima, M. H., Kojima, K., Koike-Akino, T., Jha, D., BingnanWang, and Lin, C. Deep neural network inverse design of integrated photonic power splitters. *Sci. Rep.*, 2019.

Tait, A. N., Nahmias, M. A., Shastri, B. J., et al. Broadcast and weight: An integrated network for scalable photonic spike processing. *J. Light. Technol.*, 2014.

Tait, A. N., de Lima, T. F., Zhou, E., et al. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.*, 2017.

Tang, Y., Fan, J., Li, X., Ma, J., Qi, M., Yu, C., and Gao, W. Physics-guided and physics-explainable recurrent neural network for time dynamics in optical resonances. *Nat. Compu. Sci.*, 2022.

Trivedi, R., Su, L., Lu, J., Schubert, M. F., and JelenaVuckovic. Data-driven acceleration of photonic simulations. *Sci. Rep.*, 2019.

Wang, Z., Chang, L., Wang, F., Li, T., and Gu, T. Integrated photonic metasystem for image classifications at telecommunication wavelength. *Nat Commun*, 13(1): 2131, April 2022. doi: 10.1038/s41467-022-29856-7.

Wetzstein, G., Ozcan, A., Gigan, S., Fan, S., Englund, D., Soljačić, M., Denz, C., , Miller, D. A. B., and Psaltis, D. Inference in artificial intelligence with deep optics and photonics. *Nature*, 2020.

Xu, X., Tan, M., Corcoran, B., Wu, J., Boes, A., Nguyen, T. G., Chu, S. T., Little, B. E., Hicks, D. G., Morandotti, R., Mitchell, A., and Moss, D. J. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature*, 2021.

Zhou, H., Zhao, Y., Xu, G., Wang, X., Tan, Z., Dong, J., and Zhang, X. Chip-Scale Optical Matrix Computation for PageRank Algorithm. *JSTQE*, 2020.

Zhu, H. H., Zou, J., Zhang, H., Shi, Y. Z., Luo, S. B., Wang, N., Cai, H., Wan, L. X., Wang, B., Jiang, X. D., Thompson, J., Luo, X. S., Zhou, X. H., Xiao, L. M., Huang, W., Patrick, L., Gu, M., Kwek, L. C., and Liu, A. Q. Space-efficient optical computing with an integrated chip diffractive neural network. *Nature Communications*, 13(1):1044, December 2022. doi: 10.1038/s41467-022-28702-0.