ICCAD: G: Light in Artificial Intelligence: Efficient Neurocomputing with Optical Neural Networks

Jiaqi Gu, jqgu@utexas.edu, ACM ID: 7981158

ECE Department, University of Texas at Austin, Austin, TX USA 78712

1 PROBLEM AND MOTIVATION

Deep neural networks have received an explosion of interest for their superior performance in various intelligent tasks and high impacts on our lives. The computing capacity is in an arms race with the rapidly escalating model size and data amount for intelligent information processing. Practical application scenarios, e.g., autonomous vehicles, data centers, and edge devices, have strict energy efficiency, latency, and bandwidth constraints, raising a surging need to develop more efficient computing solutions. However, as Moore's law is winding down, it becomes increasingly challenging for conventional electrical processors to support such massively parallel and energy-hungry artificial intelligence (AI) workloads. Limited clock frequency, millisecond-level latency, high heat density, and large energy consumption of CPUs, FPGAs, and GPUs motivate us to seek an alternative solution using silicon photonics. Silicon photonics is a promising hardware platform that could represent a paradigm shift in efficient AI acceleration with its CMOS-compatibility, intrinsic parallelism of optics, and near-zero power consumption. With potentially petaFLOPS per mm^2 execution speed and attojoule/MAC computational efficiency, fully-optical neural networks (ONNs) demonstrate orders-of-magnitude higher performance than their electrical counterparts [1–6]. However, previous ONN designs have a large footprint and noise robustness issues, which prevent practical applications of photonic accelerators.

In this work, we propose to explore efficient neuromorphic computing solutions with optical neural networks. Various photonic integrated circuit designs and software-hardware co-optimization methods are explored and presented here to enable high-performance photonic accelerators with lower area cost, better energy efficiency, higher variation-robustness, and more on-device learnability.

2 BACKGROUND AND RELATED WORK

Optical computing has been demonstrated as a promising substitution for electronics in efficient artificial intelligence due to its ultra-high bandwidth, sub-nanosecond latency, attojoule/MAC energy efficiency. The early research efforts focus on diffractive free-space optical computing, optical reservoir computing [7], and spike processing [8, 9] to achieve optical multi-layer perceptrons (MLPs). Recently, the integrated optical neural networks (ONNs) have attracted extensive research interests given their compactness, energy efficiency, and electronics-compatibility [1, 3, 4, 6]. Figure 1 shows the ONN design stack, including architecture and circuit design, model optimization, and final deployment and on-chip training. Due to the complexity of photonic analog circuit design and non-ideality in physical chip deployment, the power of ONNs will not be fully unleashed without careful optimization on scalability, robustness, and learnability.

In the first design stage, neural architectures and their corresponding photonic circuits will be jointly designed to map neurocomputing to optical components. Previously, Shen *et al.* [1] successfully demonstrates a singular-value-decomposition-based coherent ONN constructed by cascaded Mach-Zehnder interferometer (MZI) arrays, shown in Fig. 2(a). Their photonic tensor core prototype demonstrates

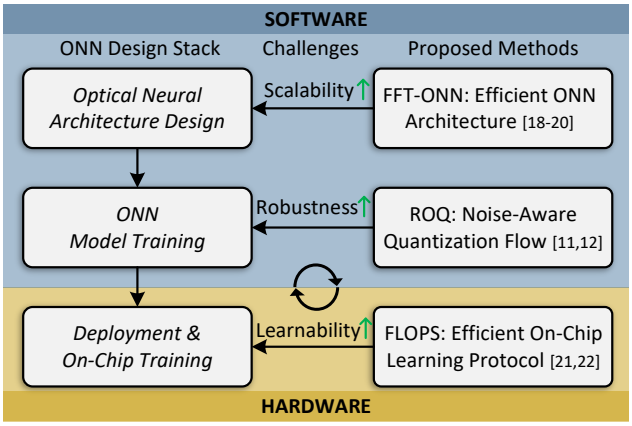
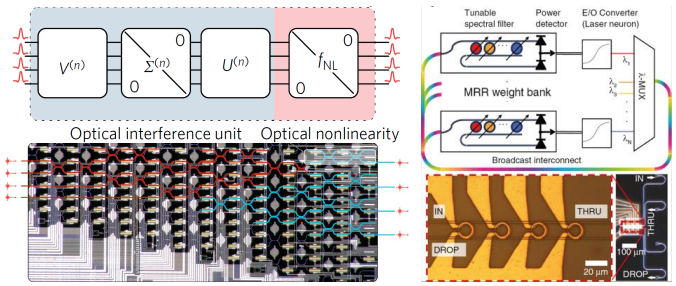


Figure 1: Challenges in current ONN design stack and the proposed hardware/software co-design solutions.



(a) (b)

Figure 2: (a) Matrix multiplication achieved by cascaded MZI arrays [1]. (b) Micro-ring resonator (MRR) based ONN [13].

order-of-magnitude higher inference throughput and comparable accuracy on the vowel recognition task compared with GPUs [1]. Zhao *et al.* [10] proposed a slimmed ONN architecture to cut down the area cost by 30-50% through a software-hardware co-design methodology. These architectures generally have a large area cost and low robustness due to phase accumulation error in the cascaded MZI meshes [11, 10, 12]. Micro-ring resonator (MRR) based incoherent ONNs have been proposed to build MRR weight banks for matrix multiplication [13, 14], shown in Fig. 2(b) the MRR-ONN has a small footprint, but it suffers from robustness issues due to MRR sensitivity. To enable scalable optical AI accelerators, novel photonic structures are in high demand to construct compact, efficient, and expressive optical neural architectures.

Given an ONN architecture, the second stage is to perform ONN model optimization. Specifically, we need to determine the optical component configurations that can achieve the AI tasks with high performance and fidelity. Previously, hardware-unaware model training is adopted to obtain a theoretically trained ONN model. However, the trained ONN weights are not necessarily implementable given limited

device control precision and non-ideal process variations [11, 12, 15]. Prior work exists to show that the error accumulation effects cause undesired sensitivity of analog ONN chips to noises, which generally lead to unacceptable accuracy drop and even complete malfunction [11, 15]. However, existing training methods lack practical estimation of device non-ideality, which leaves a large room for hardware-software co-design to bridge the performance gap between theoretical models and the physically deployed ONN chips.

The third stage happens after ONN chip manufacturing. Once there is any change in the environment, tasks, or data distribution, automatic tuning and on-chip training will be performed *in situ* to calibrate the circuit states and quickly adapt the ONN chip accordingly. Thanks to the ultra-fast execution speed and reconfigurability of silicon photonics chip, ONNs are also perfect self-learning platforms. Such self-learnability is especially beneficial to offload centralized cloud-based training to resource-limited edge devices, which not only saves expensive communication cost but also boosts the intelligence of edge computing units. To enable such on-device learnability, prior work attempts to perform *in situ* ONN training to boost the training efficiency. Brute-force device tuning [1] and evolutionary algorithms [16] are proposed to search for an optimal device configuration with a large number of ONN queries. Adjoint variable methods [17] are applied to directly generate and read out the gradients w.r.t. device configurations *in-situ* to achieve parallel on-chip backpropagation. Though the training speed is already orders-of-magnitude higher than software training, their scalability and robustness are inadequate for practical on-chip learning due to algorithmic inefficiency or prohibitive hardware overhead.

Overall, existing studies still fail to provide hardware-efficient, robust, and self-learnable ONN designs. Hence, better ONN architecture designs and advanced circuit-algorithm co-optimization are still in great demand. Therefore, we propose a holistic ONN design solution to help build scalable, reliable, and adaptive photonic accelerators with the following methodologies that can be fully integrated,

- **Frequency-Domain ONN Architecture**: *for the first time, the neural computing is mapped to a general optical frequency domain with massive parallelism.* We propose a compact and energy-efficient ONN architecture based on learnable photonic butterfly meshes. A hardware-aware structured pruning is applied to further boost the hardware efficiency by ~ 10× [18–20].
- **Nonideality-Aware ONN Optimization**: *limited device control resolution and process variations are considered during ONN optimization for the first time.* A noise-aware quantization flow is proposed to achieve considerable robustness improvement under practical circuit variations with minimum training overhead [11, 12].
- **Efficient ONN On-Chip Learning**: we propose an efficient ONN on-chip learning framework, *which is the first to enable scalable and self-learnable intelligence in integrated optics.* Our power-aware sparse zeroth-order optimization flow considerably boosts the on-device training speed by 3-8×, scalability by ~ 20×, and saves >90% training power consumption [21, 22].

As shown in Fig. 1, the proposed algorithms focus on different stages of the ONN design stack and synergistically put forward the practical application of photonic AI accelerators.

3 APPROACH AND UNIQUENESS

In this study, we propose a holistic solution to enable efficient photonic accelerator design, including a learnable frequency-domain ONN architecture `FFT-ONN` for improving area efficiency, a noise-aware quantization scheme `ROQ` for robust ONN design, and an efficient ONN on-chip learning framework `FLOPS` leveraging stochastic zeroth-order optimization for *in situ* ONN training.

3.1 `FFT-ONN`: Hardware-Efficient ONN Architecture

Though ONNs have ultra-fast execution speed, they typically have large area cost due to the physical limits of the optical devices. Previous MZI-based ONNs [1, 10] consume a large number of MZIs, thus fail to provide efficient and scalable photonic solutions. We focus on answering the following critical questions: 1) how to construct a new ONN architecture using much fewer optical components without sacrificing its expressiveness, 2) how to efficiently map high-dimensional neural networks to 2-dimensional (2D) photonic integrated chips (PICs), and 3) how to further cut down the power consumption of optical devices with minimum performance degradation [18–20].

Proposed Frequency-Domain Optical MLP: To remedy the area cost issue, we propose a novel optical multi-layer perceptron architecture based on fast Fourier transform, as shown in Fig. 3 Instead of implementing general matrix multiplication, we adopt a block-circulant weight matrix as an efficient substitution. This block-circulant matrix can efficiently realize restricted linear projection via an FFT-based fast multiplication algorithm. After transformed into the Fourier domain, matrix multiplication can be cast to lightweight element-wise multiplication between two Fourier-domain signals. A photonic butterfly network is designed using compact photonic devices to achieve on-chip optical FFT/IFFT. Frequency-domain weights are implemented in the element-wise multiplication (EM) stage to achieve complex-valued multiplication by leveraging the polarization feature of light. Optical splitter and combiner trees are designed to perform light-speed fan-out and partial product accumulation, respectively. This framework has provably comparable model expressivity with classical NNs. Without sacrificing any classification accuracy, this frequency-domain optical multi-layer perceptron [18] demonstrated 3-4× lower area cost compared with previous ONNs [1, 10].

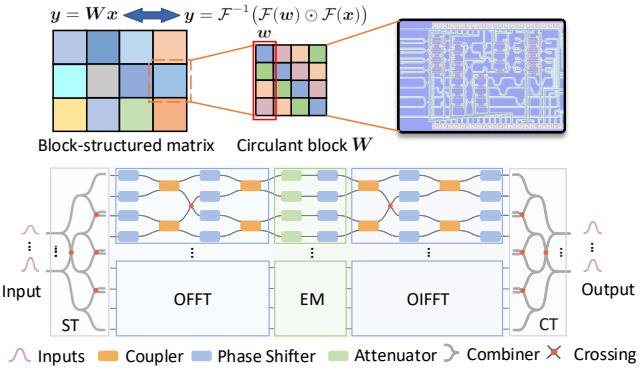


Figure 3: `FFT-ONN` [18] architecture and ONN chip tape-out.

Proposed Frequency-Domain Optical CNN: To support highly parallel 2D convolutional neural network (CNN) acceleration, we propose a learnable frequency-domain optical CNN architecture `FFT-ONN-v2`, which moves beyond the traditional FFT-based design methodology and fundamentally differs from traditional `im2col`-based spatial CNN accelerators. To match high-dimensional convolutions to 2D photonic circuits, we map the convolution to *spatial, temporal, and spectral* dimensions of the PIC, fully unleashing the massive parallelism of ONNs. We decompose the 2D spatial convolution into two

cascaded 1D frequency-domain convolutions along rows and columns. In the column-wise convolution, the feature maps are projected to the frequency domain via learnable butterfly transform structures column by column, multiplied by the frequency-domain kernels, and projected again by the reversed transform. We use multiple ultra-compact micro-disk (MD) weight banks to directly implement all complex-valued kernels *at one shot*. We allow an augmented solution space for better model expressivity without conjugate symmetry constraints in the traditional FFT-based method. Wavelength-division multiplexing (WDM) techniques are adopted to allow extensive hardware reuse and massive parallelism across channels, leading to order-of-magnitude area reduction and throughput improvement.

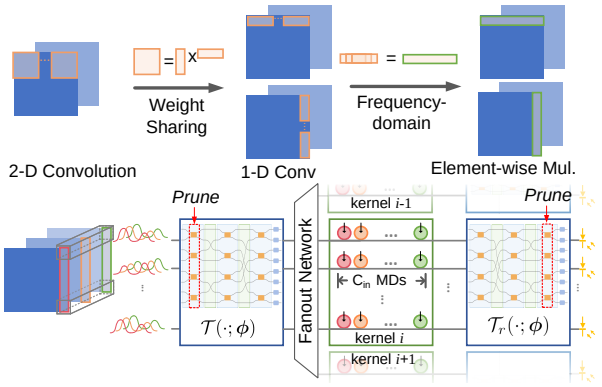


Figure 4: FFT-ONN-v2 [19] on frequency-domain CNNs.

Proposed Power-Aware Optimization: The manually designed optical FFT module is a fixed structure with limited reconfigurability. It can be sub-optimal in terms of area and power consumption. To empower our ONN with joint learnability, the *fixed* FFT module is relaxed into a *general* frequency-domain projection with a trainable butterfly structure. All programmable devices in the transform are jointly optimized during training to learn the best transform pairs automatically. We also employ hardware-aware fine-grained structured pruning and progressively sparsify the phase shifters in the learnable transform to cut down power consumption and the circuit footprint.

Overall Contributions: We propose a hardware-efficient ONN architecture to break through the ONN scalability and efficiency. It is the first time that neural computing has been mapped to a general frequency domain for integrated optics. We move beyond the traditional FFT-based NN design paradigm and propose an ultra-compact frequency-domain ONN architecture with algorithmic and hardware insights. We also adopt a hardware-aware pruning method to further improve its area and power. Our compact and highly parallel FFT-ONN-family achieves ~10× area reduction and 10× device-tuning power saving compared with previous ONNs.

3.2 ROQ: Noise-Aware Quantization Scheme for Robust ONNs

As analog computing platforms, ONNs inevitably encounter robustness issues due to process variations and a non-ideal environment. The limited device control resolution, e.g., typically 4-bit, is another practical factor that potentially induces undesired accuracy degradation. However, prior ONN model optimization methods lack effective noise modeling, such that the deployed ONN model suffers from severe accuracy drop or even complete malfunction [1, 11, 15]. Instead of hardware-unaware ONN optimization, we focus on two aspects: 1) how to find quantized device configurations that honor control resolution limits while maintaining model expressivity, and 2) how to efficiently consider device variations into training to find a more noise-resilient solution via hardware-software co-design [11, 12].

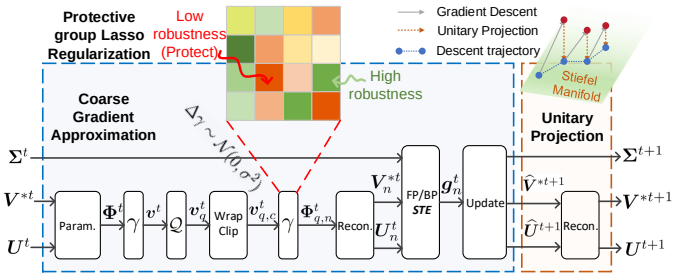


Figure 5: ROQ [12] for robust ONNs under low-bit controls.

Proposed Differentiable ONN Quantization Flow: We target the general photonic tensor core design based on MZI arrays [1], the most challenging design without readily applicable prior quantization methods. MZI mesh simulation will be performed for hardware-aware training to emulate the impact of low control resolution effects on the unitary transfer matrix. The biggest challenge is the prohibitive cost of propagating gradients through the discretized phase decomposition [1]. To efficiently tackle the optimization challenge for ONNs, we propose a differentiable phase-domain quantization scheme, shown in Fig. 5. In our proposed coarse gradient approximation algorithm, the training engine directly bypasses the upstream gradients through the entire discretized decomposition procedure. Then, we can efficiently estimate the gradients w.r.t the unitary matrices. Each gradient descent step will push the unitary matrices out of the unitary manifold, which is illegal for the subsequent unitary decomposition. We adopt a projected gradient descent optimizer to enable efficient optimization in the discretized unitary space.

Proposed Protective Noise Injection: Optical devices inevitably have non-ideal parameter drifts due to process variation and manufacturing error. To model this circuit noise in our quantization flow, we propose a protective Group Lasso (PGL) regularization technique to explicitly perform noise-adaptive optimization. Random device drifts estimated from foundry PDKs will be injected to emulate the variation. A block-wise robustness estimator will be used to assign a robustness score for each block based on the induced errors. Highly sensitive weight blocks will be suppressed to protect the ONN from accumulated errors.

Overall Contributions: Our noise-aware quantization scheme introduces coarse gradient approximation and unitary projection algorithms to enable differentiable ONN optimization with non-ideality modeling. Our protective noise injection method efficiently considers device noise modeling during training to improve the ONN noise tolerance. Our proposed algorithm-circuit co-optimization methodology ROQ shows much better noise tolerance under low-bit device controls and practical circuit variations, enabling general photonic tensor accelerators with low control complexity and high robustness.

3.3 FLOPS: Efficient On-Chip ONN Learning

Besides inference acceleration, efficient on-device training is another critical step in intelligent edge AI solutions, which requires efficient learning protocols to be developed, especially for resource-limited edge devices. However, traditional software-based ONN training suffers the problems of expensive hardware mapping and inaccurate variation modeling. Previous on-chip learning protocols are either based on brute-force tuning [1] or evolutionary algorithms [16], which

fail to leverage the self-learning capability of ONNs due to algorithmic inefficiency and poor variation-robustness. To enable ultra-fast training acceleration on self-learnable photonic neural engines, we propose an efficient on-chip learning framework FLOPS [21, 22] to resolve the scalability challenges using stochastic zeroth-order optimization.

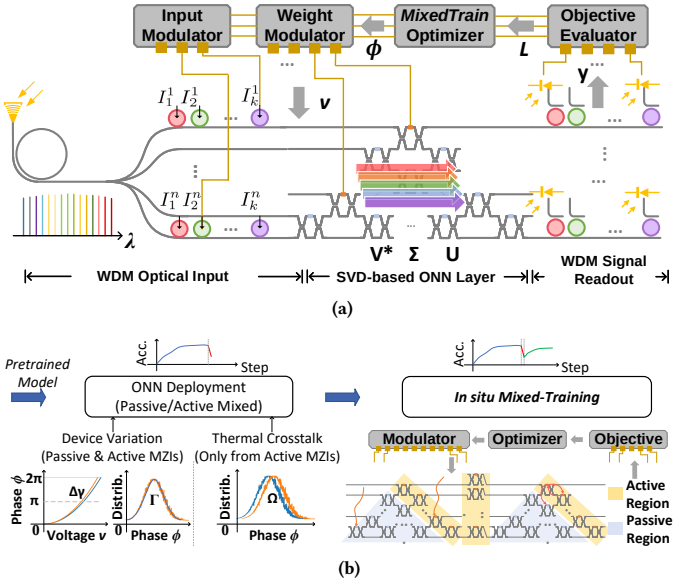


Figure 6: (a) ONN on-chip training framework and (b) sparse mixed-training strategy [21, 22].

Proposed Zeroth-Order Learning Framework: As shown in Fig. 6(a), ONN chips are naturally ultra-fast DNN forward accelerators. Hence, we directly train optical devices on chip to achieve efficient *in situ* optimization without costly gradient backpropagation. We propose a stochastic zeroth-order gradient estimator to efficiently query the ONN oracle for variance-reduced gradient descent. WDM techniques are utilized to enable fully parallel evaluation on a mini-batch of training data, which considerably reduces the ONN query complexity. By leveraging the ONN chip itself as an accurate *in-situ* variation model, we can perform on-device training without expensive noise simulation to efficiently recover the accuracy with high noise robustness under post-deployment non-ideality.

Proposed Sparse Mixed-Training Strategy: On resource-constrained edge platforms, critical barriers for on-device learnability are energy budget and resource limits. We propose a mixed-training strategy with two-level sparsity to improve the ONN training efficiency, shown in Fig. 6(b). We partition the optical devices into passive and active regions and only allow active devices to be trainable. Such parameter-level sparsity considerably reduces the device programming power and the inter-device crosstalk without degrading the model reconfigurability. During each optimization step, we also explore gradient-level sparsity by randomly selecting a small subset of devices to update their configurations. We further apply lightweight dynamic pruning to explicitly optimize learning energy cost by randomly removing exploration steps with extra power cost, leading to order-of-magnitude training power reduction without accuracy drop.

Overall Contributions: With the proposed on-chip mixed-training framework applied, we can outperform previous on-chip learning methods with 3-8× faster training speed, 3-5% higher robustness, >20× better scalability, and over 90% power reduction.

4 RESULTS AND CONTRIBUTIONS

4.1 FFT-ONN: Hardware-Efficient ONN Architecture

To validate the functionality and efficiency of our proposed ONN architecture, we first perform optical simulation using commercial tools Lumerical INTERCONNECT. Our model shows good fidelity with <1.2% max relative error compared with theoretical results.

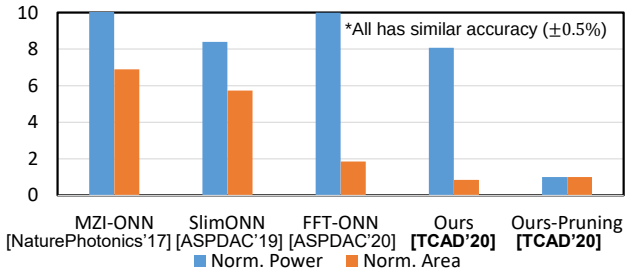


Figure 7: Area and power comparison with prior ONNs [1, 10].

We further compare the accuracy and hardware cost with prior state-of-the-art (SOTA) ONNs [1, 10]. On several optical MLP and CNNs with various benchmarks, e.g., MNIST and FashionMNIST, our FFT-based optical MLP can save 2.2×-3.7× device usage compared with prior ONNs [1, 10]. With our learnable transform and hardware-aware pruning, the frequency-domain ONN architecture considerably boosts the area efficiency by nearly 7× with 80-90% device programming power reduction. We also compare the noise-robustness under practical device variations. Our pruned butterfly structure demonstrates superior robustness with much fewer noise sources. With 80% structured sparsity, FFT-ONN-v2 maintains over 97% accuracy on MNIST, while prior MZI-ONNs suffer from complete malfunction.

4.2 ROQ: Noise-Aware Quantization Scheme for Robust ONNs

We evaluate the effectiveness of our proposed quantization scheme on a four-layer ONN with the MNIST dataset and compare it with naive hardware-unaware training and a baseline iterative quantization method. Based on a pre-trained full-precision ONN model with 97.6% accuracy, we quantize the device control signals with 3- to 6-bit precision and inject practical process variations.

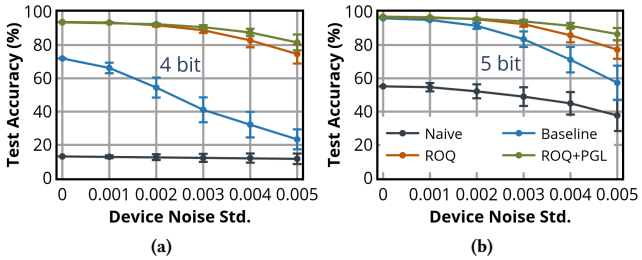


Figure 8: Robustness comparison between the proposed scheme ROQ [12] and baseline methods.

Figure 8 shows that our proposed ROQ and protective regularization technique (PGL) achieve the highest accuracy with the lowest variances on all settings. With large device noises and only 3-bit resolution, our method can boost the accuracy from ~20% (baseline) to 80%. Our proposed ROQ effectively tackles the non-ideal issues of

ONNs via co-design and provides a low-overhead model optimization approach towards noise-resilient ONN accelerators.

4.3 FLOPS: Efficient On-Chip ONN Learning

We compare the training efficiency with prior SOTA on-chip training protocols in terms of the number of ONN forward, recovered accuracy, and learning power consumption.

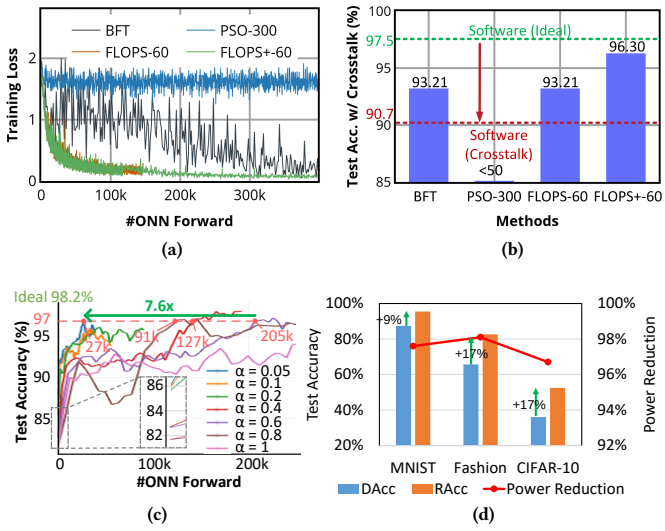


Figure 9: (a) Learning curve comparison [21]. (b) Accuracy under device variations [21]. (c) Learning curve with different mixed-training sparsity α [22]. (d) Deployed accuracy (DAcc), Recovered (RAcc) accuracy, and power reduction [22].

Figure 9(a) and 9(b) shows that our FLOPS achieves 3-4× higher learning efficiency with 3% higher robustness than prior arts [1, 16]. With our two-level sparse mixed-training strategy, the learning efficiency is boosted by >7× compared with the baseline method [1], shown in Fig. 9(c). Our optimizer can handle the ONNs training with 2500 MZIs, showing >20× higher scalability than prior protocols [1, 16]. With our dynamic power optimization, 96%-98% training power is saved compared to our DAC-version [21], with marginal overhead and negligible accuracy drop, shown in Fig. 9(d). Our proposed on-chip learning solution enables scalable and fast on-device training to facilitate intelligent and adaptive photonic accelerators.

4.4 Research Impacts

As the major focus of my Ph.D. researches, this study leads to **7 first-authored publications** [18, 12, 21, 19, 23, 24, 22] in premier EDA/-CAD/ML journals and conference, such as TCAD, ASP-DAC, DATE, DAC, and AAAI. In addition, this study leads to **2 invited paper** [11] at ICCAD 2019 [11], CLEO 2021 [25], and **11 co-authored high-impact SPIE/Nature journals and conferences** [2, 20, 26–28]. The proposed design methodologies are built on advanced machine learning algorithms and solid physical optical modeling and facilitate the entire ONN design flow. We pioneer the research in the area, and our proposed FFT-ONN [18] received the **Best Paper Award** in ASP-DAC 2020. The proposed FLOPS [21] was selected as one out of 6 **Best Paper Finalists** at DAC 2020. Our endeavor on optics-AI integration is well recognized by the academia, and we receive the **Gold Medal** in ACM/SIGDA student research competition and the **Best Poster Award** at NSF Machine Learning Hardware Workshop 2020. Our ONN designs have a photonic neural chip tape-out with AMF for measurement and prototype demonstration.

REFERENCES

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

[2] Z. Ying, C. Feng, Z. Zhao, S. Dhar, H. Dalir, J. Gu, Y. Cheng, R. Soref, D. Z. Pan, and R. T. Chen, "Electronic-photonic arithmetic logic unit for high-speed computing," *Nature Communications*, 2020.

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

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

[5] M. A. Nahmias, T. F. de Lima, A. N. Tait, H. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic multiply-accumulate operations for neural networks," *JSTQE*, 2020.

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

[7] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nature Communications*, 2013.

[8] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: An integrated network for scalable photonic spike processing," *J. Light. Technol.*, 2014.

[9] D. Rosenbluth, K. Kravtsov, M. P. Fok *et al.*, "A high performance photonic pulse processing device," *Opt. Express*, vol. 17, no. 25, Dec 2009.

[10] Z. Zhao, D. Liu, M. Li, Z. Ying, L. Zhang, B. Xu, B. Yu, R. T. Chen, and D. Z. Pan, "Hardware-software co-design of slimmed optical neural networks," in *Proc. ASPDAC*, 2019.

[11] Z. Zhao, J. Gu, Z. Ying, C. Feng, R. T. Chen, and D. Z. Pan, "Design technology for scalable and robust photonic integrated circuits," in *Proc. ICCAD*, 2019.

[12] J. Gu, Z. Zhao, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "ROQ: A noise-aware quantization scheme towards robust optical neural networks with low-bit controls," in *Proc. DATE*, 2020.

[13] A. N. Tait, T. F. de Lima, E. Zhou, , A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, 2017.

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

[15] Y. Zhu, G. L. Zhang, B. Li, X. Yin, C. Zhuo, H. Gu, T. Y. Ho, and U. Schlichtmann, "Countering Variations and Thermal Effects for Accurate Optical Neural Networks," in *Proc. ICCAD*, 2020.

[16] T. Zhang, J. Wang, Y. Dan, Y. Lanqiu, J. Dai, X. Han, X. Sun, and K. Xu, "Efficient training and design of photonic neural network through neuroevolution," *Optics Express*, 2019.

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

[18] J. Gu, Z. Zhao, C. Feng, M. Liu, R. T. Chen, and D. Z. Pan, "Towards area-efficient optical neural networks: an FFT-based architecture," in *Proc. ASPDAC*, 2020.

[19] J. Gu, Z. Zhao, C. Feng, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability," *IEEE TCAD*, 2020.

[20] C. Feng, J. Gu, Z. Ying, Z. Zhao, R. T. Chen, and D. Z. Pan, "Scalable fast-Fourier-transform-based (FFT-based) integrated optical neural network for compact and energy-efficient deep learning," in *SPIE Photonics West*, 2021.

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

[22] J. Gu, C. Feng, Z. Zhao, Z. Ying, R. T. Chen, and D. Z. Pan, "Efficient On-Chip Learning for Optical Neural Networks Through Power-Aware Sparse Zeroth-Order Optimization," in *Proc. AAAI*, 2021.

[23] J. Gu, Z. Zhao, C. Feng, Z. Ying, R. T. Chen, and D. Z. Pan, "O2NN: Optical Neural Networks with Differential Detection-Enabled Optical Operands," in *Proc. DATE*, Feb. 2021.

[24] J. Gu, C. Feng, Z. Zhao, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators," in *Proc. DATE*, Feb. 2021.

[25] J. Midkiff, A. Rostamian, K. M. Yoo, , A. Asghari, C. Wang, C. Feng, Z. Ying, J. Gu, H. Mei, C.-W. Chang, J. Fang, A. Huang, J.-D. Shin, X. Xu, M. Bukshtab, D. Z. Pan, and R. T. Chen, "Integrated Photonics for Computing, Interconnects and Sensing," in *Proc. CLEO*, May 2021.

[26] M. Miscuglio, Z. Hu, S. Li, J. Gu, A. Babakhani, P. Gupta, C.-W. Wong, D. Pan, S. Bank, H. Dalir, and V. J. Sorger, "Million-channel parallelism Fourier-optic convolutional filter and neural network processor," in *Proc. CLEO*, 2020.

[27] C. Feng, Z. Zhao, Z. Ying, J. Gu, D. Z. Pan, and R. T. Chen, "Compact design of On-chip Elman Optical Recurrent Neural Network," in *Proc. CLEO*, 2020.

[28] C. Feng, Z. Ying, Z. Zhao, J. Gu, D. Z. Pan, and R. T. Chen, "Wavelength-division-multiplexing (WDM)-based integrated electronic–photonic switching network (EPSN) for high-speed data processing and transportation," *Nanophotonics*, 2020.