Characterization and Optimization of Coherent MZI-Based Nanophotonic Neural Networks Under Fabrication Non-Uniformity

Asif Mirza, *Student Member, IEEE*, Amin Shafiee, Sanmitra Banerjee, *Student Member, IEEE*, Krishnendu Chakrabarty, *Fellow, IEEE*, Sudeep Pasricha, *Senior Member, IEEE*, and Mahdi Nikdast, *Senior Member, IEEE*

***Abstract*—Silicon-photonic neural networks (SPNNs) and artificial intelligence (AI) accelerators have emerged as promising successors to electronic accelerators by offering orders of magnitude lower latency and higher energy efficiency. Nevertheless, the underlying silicon photonic devices in SPNNs are sensitive to inevitable fabrication-process variations (FPVs) stemming from optical lithography imperfections. Consequently, the inferencing accuracy in an SPNN can be highly impacted by FPVs—e.g., can drop to below 10%—the impact of which is yet to be fully studied. In this article, we, for the first time, model and explore the impact of optical phase noise due to FPVs in the waveguide width, silicon-on-insulator (SOI) thickness, and etch depth in coherent SPNNs that use Mach–Zehnder interferometers (MZIs). Leveraging such models, we propose a novel variation-aware, design-time optimization solution to improve MZI tolerance to different FPVs in SPNNs. Simulation results for two example SPNNs of different scales under realistic and correlated FPVs indicate that using the optimized MZIs can lead to significant improvements in the network inferencing accuracy. The proposed one-time optimization imposes low area overhead and hence is applicable even to resource-constrained designs.**

***Index Terms*—Silicon photonic integrated circuits, fabrication-process variations, deep learning, optical neural networks.**

I. Introduction

MACHINE learning algorithms are being utilized in a wide range of applications ranging from autonomous driving, real-time speech translation, and network anomaly detection to pandemic growth and trend prediction. With the rising demand for advanced neural networks to address even more complex problems, artificial intelligence (AI) accelerators need to consistently deliver better performance and improved accuracy while being energy-efficient. Unfortunately, in the post-Moore's law era, electronic AI accelerator architectures face fundamental limits in their processing capabilities due to the limited bandwidth and low energy efficiency of metallic interconnects. Silicon photonics can alleviate these bottlenecks by enabling chip-scale optical interconnects with ultra-high bandwidth, low-latency, and energy-efficient communication, and light-speed chip-scale optical computation [1]. Leveraging silicon-photonic-enabled optical interconnect and computation, many integrated silicon-photonic neural networks (SPNNs) have been recently proposed [1].

Prior work has shown that the complexity of matrix-vector multiplication can be reduced from $O(N^2)$ to $O(1)$ in SPNNs [2]. Also, as SPNNs use photons for computation, there is negligible latency associated with inferencing. Such benefits have positioned SPNNs as a promising alternative to traditional electronic neural networks [3]. Nevertheless, fabrication-process variations (FPVs) due to inevitable optical lithography imperfections lead to undesired changes in the critical dimensions of SPNNs' building blocks (e.g., waveguide width and thickness in Mach–Zehnder interferometers (MZIs) in coherent SPNNs), imposing inaccuracies during matrix-vector multiplication. In [4], we have shown that random uncertainties due to FPVs and thermal crosstalk can result in up to a catastrophic 70% accuracy loss, even in mature fabrication processes. Existing approaches for improving the resilience of SPNNs against FPVs largely rely on compensating their impact by individually calibrating MZIs in the network [5], [6]. However, such solutions impose additional calibration (i.e., tuning) power overhead and are not scalable as their complexity grows with the number of devices in SPNNs.

The main contribution of this article is in developing, to the best of our knowledge, the first comprehensive analysis of the impact of physical-level FPVs on coherent SPNN performance. We consider variations in the waveguide width, silicon-on-insulator (SOI) thickness, and etch depth, which impacts the slab thickness, based on realistic FPV maps developed using different correlation lengths in the variations and experimental

Manuscript received 11 September 2022; accepted 8 November 2022. Date of publication 23 November 2022; date of current version 7 December 2022. This work was supported by the National Science Foundation (NSF) under Grants CCF-1813370, CCF-2006788, and CNS-2046226. The review of this article was arranged by guest editors of the Special Issue for NanoCoCoA2021. *(Asif Mirza, Amin Shafiee, and Sanmitra Banerjee contributed equally to this work.) (Corresponding author: Mahdi Nikdast.)*

Asif Mirza, Amin Shafiee, Sudeep Pasricha, and Mahdi Nikdast are with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523 USA (e-mail: mirza.baig@colostate.edu; amin.shafiee@colostate.edu; sudeep@colostate.edu; mahdi.nikdast@colostate.edu).

Sanmitra Banerjee is with the NVIDIA Corporation, Santa Clara, CA 95051 USA (e-mail: sanmitra.banerjee@duke.edu).

Krishnendu Chakrabarty is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: krish@duke.edu).

Digital Object Identifier 10.1109/TNANO.2022.3223915

measurements. We model the impact of FPVs at the device level for MZI performance and at the network level for arrays of cascaded MZIs in SPNNs. At the system level, we explore the impact of variations on SPNN inferencing accuracy while considering different FPVs and variation correlation lengths. Leveraging our detailed device-level models, we also develop a novel design optimization solution to improve MZI performance in SPNNs under different FPVs. Our simulation results for two example SPNNs (with 1380 and 20,580 phase shifters) under FPVs show that while inferencing accuracy can drop to as low as 7.73% under different variations, using our optimized MZIs can significantly improve the inferencing accuracy (e.g., by up to 72% on average in a large SPNN). Note that this article does not consider the impact of thermal crosstalk and FPVs in directional couplers in MZIs.

The rest of the article is organised as follows. Section II presents a background on MZIs and SPNNs and a summary of prior related work. In Section III, we analyze the impact of FPVs on MZIs (device level) and array of cascaded MZIs (network level) in SPNNs. Section IV presents our MZI design optimization solution to design devices with improved tolerance to different FPVs. Section V presents simulation results that highlight how the optimized MZIs can improve the accuracy of the unitary transformation (at the layer level) and the inferencing accuracy (at the system level), in the presence of FPVs. Last, we draw conclusions in Section VI.

II. Background and Prior Related Work

This section summarizes fundamentals of MZIs and coherent SPNNs designed based on MZIs. Also, it presents some preliminary models to help understand the impact of FPVs on photonic waveguides, and reviews some prior related work.

A. Mach–Zehnder Interferometer (MZI)

A 2×2 MZI in an SPNN consists of two tunable phase shifters (ϕ and θ) on the upper arm and two 50:50 beam splitters as shown in Fig. 1(c). The phase shifters are used to apply configurable phase shifts and obtain varying degrees of interference between the optical signals traversing the MZI arms. In SPNNs, phase shifters are often implemented using thermal micro-heaters for lower optical loss, where the effective index of the underlying waveguide changes with temperature (i.e., due to thermo-optic effect of silicon), hence altering the phase of the optical signal. Note that the proposed analyses in this article are independent of the phase-shift mechanism in the MZI. Moreover, 2×2 beam splitters can be designed using directional couplers (DCs) where a fraction of the optical signal (defined as κ) at an input port is transmitted to an output port, and the remaining signal is coupled to the other output port, as shown in the inset of Fig. 1(c). The ideal transfer matrix (T_{MZI}) for an MZI with two phase shifters (θ and ϕ) and two 50:50 beam splitters can be defined as [7] (Fig. 1(c)):

$$T_{MZI}(\theta,\phi) = \begin{pmatrix} \frac{e^{i\phi}}{2}\left(e^{i\theta}-1\right) & \frac{i}{2}\left(e^{i\theta}+1\right) \\ \frac{ie^{i\phi}}{2}\left(e^{i\theta}+1\right) & -\frac{1}{2}\left(e^{i\theta}-1\right) \end{pmatrix}. \quad (1)$$

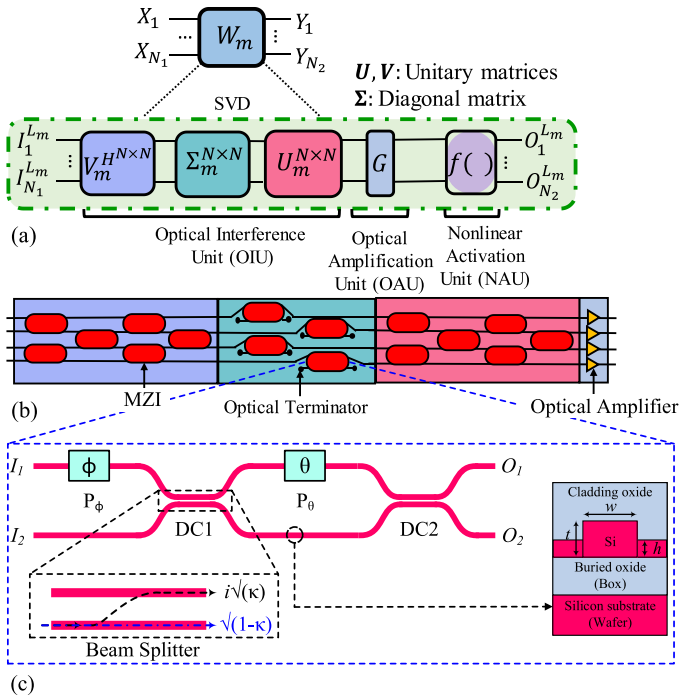


Fig. 1. (a) Overview of singular value decomposition (SVD) of a weight matrix related to a fully connected layer (L_m) with N_1 as the number of input ports and N_2 as the number of output ports. (b) An optical-interference unit (OIU). (c) A 2×2 MZI structure with two integrated phase shifters (θ and ϕ) and two beam splitters based on directional couplers (DCs).

B. Coherent SPNNs Based on MZIs

Compared to non-coherent SPNNs, coherent SPNNs—considered in this paper—use a single wavelength and MZI devices, in which the adjusted phase shifts in the phase shifters denote the dynamic weight parameters. Fig. 1(a) shows an example of a coherent SPNN. A fully connected layer in a deep neural network (L_m) can be realized with n_m neurons. Each layer performs a linear matrix-vector multiplication and accumulation (MAC) and passes the outputs to the next layer. The output vector of L_m can be mathematically modeled as $O_m^{n_m\times 1} = f_m(W_m \times O_{m-1}^{n_{m-1}\times 1})$. Here, f_m is a non-linear activation function (performed by non-linear activation unit (NAU) in Fig. 1(a)) of L_m, and W_m is the corresponding weight matrix of L_m. Given a weight matrix W_m, which can be obtained by training the network and mapped to MZIs using singular value decomposition (SVD), each weight matrix W_m can be written as $W_m = U_m^{n_m\times n_m}\Sigma_m^{n_m\times n_m}V_m^{H,n_m\times n_m}$. Here, $U_m^{n_m\times n_m}$ and $V_m^{H,n_m\times n_m}$ are the unitary matrices with dimension of $n_m \times n_m$, and $\Sigma_m^{n_m\times n_m}$ is a diagonal matrix with dimension of $n_m \times n_m$. Also, $V_m^{H,n_m\times n_m}$ is the Hermitian-transpose of $V_m^{n_m\times n_m}$. A unitary matrix can be realized by using a cascaded array of 2×2 MZIs. As shown in Fig. 1(a) and (b), this unit is called the optical-interference unit (OIU) and is responsible for performing the MAC operation.

Several approaches have been proposed to design the architecture of MZI arrays to perform MAC operations in the optical domain [8], [9], [10]. Out of these, the Clements design, due its symmetric nature, has a low optical loss and footprint. Therefore,

we use the Clements design [9] to transform each unitary matrix ($U_m^{n_m \times n_m}$ and $V_m^{H,n_m \times n_m}$) to a cascaded MZI array with a specific phase setting per MZI in the network. In the Clements method, each unitary matrix will be mapped to a cascaded MZI array with a total number of $\frac{N(N-1)}{2}$ MZIs, where N is the size of the designated unitary matrix. Note that the diagonal matrix ($\Sigma_m^{n_m \times n_m}$) can be realized with an array of MZIs with one input and one output terminated. The optical-amplification unit (OAU) in Fig. 1(a), which is required to obtain arbitrary diagonal matrix, can be realized using semiconductor optical amplifiers (SOAs).

C. Fabrication-Process Variations (FPVs)

FPVs in silicon photonics originate in optical-lithography process imperfections, contributing to changes in the waveguide width, SOI thickness (dominated by the host wafer), and slab thickness (in case of a ridge waveguide). Such changes deviate the effective index (n_{eff}) in a waveguide and in turn the propagation constant (β) which determines the optical phase of the signal traversing the waveguide. Considering Fig. 1(c) and as an example, the effective index (n_{eff}) in a ridge waveguide depends on the optical wavelength and the critical dimensions of the waveguide [11], i.e., width (w), SOI thickness (t), and slab thickness (h) in Fig. 1(c). Note that $h = 0$ for a strip waveguide. The relation can be defined as:

$$n_{eff}(\lambda, w, t, h) = \left(\frac{\lambda}{2\pi}\right)\beta(\lambda, w, t, h), \tag{2}$$

where λ is the optical wavelength. Here, β can be defined as $\beta = \psi/L$, where ψ is the single pass phase-shift induced in a waveguide of length L. Leveraging (2), propagation constant changes ($\Delta\beta$) in a ridge waveguide under FPVs is given by:

$$\Delta\beta = \frac{2\pi}{\lambda}\left(\frac{\partial n_{eff}}{\partial w}\rho_w + \frac{\partial n_{eff}}{\partial t}\rho_t + \frac{\partial n_{eff}}{\partial h}\rho_h\right). \tag{3}$$

Here, ρ_w, ρ_t, and ρ_h are the variations in the waveguide width, SOI thickness, and slab thickness, respectively. When using a strip waveguide, $\frac{\partial n_{eff}}{\partial h} = \rho_h = 0$. Note that, in this article, we do not consider the variations in L (see Section IV).

D. Related Work on FPV Analysis in SPNNs

Our prior work in [4] studied the impact of random phase noise due to FPVs and thermal crosstalk at the system level in SPNNs by developing a framework that identifies critical components in the network. In [7], imprecisions were introduced in SPNNs after software training such that pre-fabrication training tends to be more scalable in terms of network size and volume. This helps designing precise and cost effective MZIs, when compared to re-configurable ones, which can be exploited for AI applications to perform matrix multiplication. A method was presented in [5] to counter the impact of both FPVs and thermal effects using modified cost functions during training with added benefits of post-fabrication hardware calibration. The impact of FPVs can also be reduced by minimizing the tuned phase angles in an SPNN; this can be done by leveraging the non-uniqueness of SVD as it was shown in [12]. The work in [6] proposed a circuit-level hardware error correction solution for SPNNs in which local error correction requires characterization of each phase shifter and passive splitter in the photonic circuit while relying on detectors in the output to calibrate the parameters.

TABLE I
PARAMETERS USED TO GENERATE FPV MAPS

Design Parameter	Correlation Length (l)	Standard Deviation (σ)
Waveguide width	1 mm and 100 μm	$\sigma_w = 5$ nm
SOI thickness	1 mm and 100 μm	$\sigma_t = 2$ nm
Slab thickness	1 mm and 100 μm	$\sigma_h = 2.5$ nm

The aforementioned methods focus on mitigating deviations in SPNNs *post-fabrication* by either post-fabrication training methods to compensate for any additionally introduced phase noises, which might impact the network accuracy [5], or calibrating each noisy component, where additional error-detection phases are required and the complexity can increase as the SPNN scales up [6]. In this work, we focus on exploring and optimizing the physical design of MZI devices in coherent SPNNs under FPVs *prior to fabrication*, hence improving the network tolerance under different FPVs. We show that by exploring and optimizing MZI device physical-level design, we can improve the relative-variation distance (RVD) in SPNNs, which quantifies the deviation between the intended unitary matrix and the deviated unitary matrix, to enhance the overall inferencing accuracy.

III. MODELING FPVS IN COHERENT SPNNS

This section presents a detailed bottom-up analysis of the impact of FPVs in the waveguide width, SOI thickness, and slab thickness at the device level (i.e., MZI devices) and network level (OIU in Fig. 1(b)) in coherent SPNNs. We show the impact of FPVs on SPNN inferencing accuracy (system level) in Section V.

A. Device Level: MZI Performance Under FPVs

To study the impact of FPVs on MZI devices, we should first model FPVs in silicon photonics and explore how MZI devices experience such FPVs. In our prior work [13], we have developed realistic wafer variation maps that model radial-variation effects and correlation among different variations—both of which are critical to realistically model FPVs in silicon photonics—in SOI wafers. Such maps were developed based on mean, standard deviation (σ), and correlation lengths (l) experimentally characterized in collaboration with CEA-Leti in [13]. Table I summarizes different parameters considered to generate FPV maps for our calculations, which are based on analyzing experimental data from characterizing actual 200-mm wafers at CEA-Leti. Fig. 2-top shows examples of wafer and die maps generated using our in-house FPV wafer map simulator with a resolution of 10 μm $\times$10 μm (i.e., the mesh size in the map—see Fig. 2).

MZIs are bulky devices (e.g., $\approx$340 μm in length [14]), and hence every section of the device will experience slightly different FPVs (see Fig. 2). Such a difference changes with the correlation length in the variations (e.g., long-range versus short-range correlated variations). As a result, it is critical to analyze the impact of the non-uniformity of FPVs in MZI

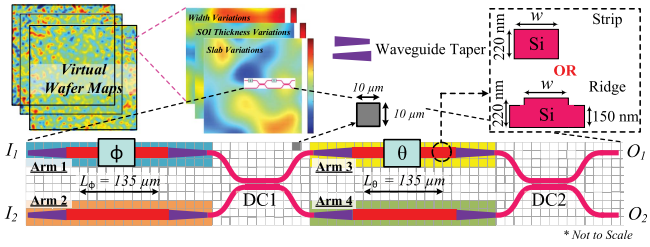


Fig. 2. An MZI device structure with waveguide tapers mapped to FPV maps (top), based on [13], with a mesh size of 10 μm. The MZI can use strip waveguides or ridge waveguides, both with the SOI thickens of 220 nm and varying waveguide width (w) on each arm. The design of slab thickness (150 nm) in the ridge waveguide is discussed in Section IV. Note that variation-free directional couplers (DCs) are considered.

devices. Considering the MZI shown in Fig. 2 with its four arms labeled (Arm1–Arm4), we average the width, SOI thickness, and slab thickness (when using a ridge waveguide in the MZI) variations observed on each MZI arm, separately over the section colored on each arm.

Considering the MZI in Fig. 2, the optical signals traversing the two opposite arms of the device—before the input DC (i.e., Arms 1 and 2) and output DC (i.e., Arms 3 and 4)—should only experience the desired phase change ϕ or θ, adjusted after the network training. Note that the phase shifters are integrated on the top of the silicon waveguides. However, due to the impact of non-uniform FPVs on each individual arm, the optical signals experience some undesired phase changes. Assuming that each arm's length ($L_1 = L_2 = L_\phi$ and $L_3 = L_4 = L_\theta$) does not undergo any variations, the optical phase difference between the two optical signals traversing the opposite arms (Arms 1–2 and Arms 3–4) and interfering at the input of DC1 ($\Delta\Phi_{DC1}$) and DC2 ($\Delta\Phi_{DC2}$) is:

$$\Delta\Phi_{DC1} = \phi + \left|\Delta\beta_1 L_1 - \Delta\beta_2 L_2\right|, \tag{4a}$$

$$\Delta\Phi_{DC2} = \theta + \left|\Delta\beta_3 L_3 - \Delta\beta_4 L_4\right|. \tag{4b}$$

Here, $\Delta\beta_1$, $\Delta\beta_2$, $\Delta\beta_3$, and $\Delta\beta_4$ are, respectively, the propagation constant changes on MZI's Arms 1–4, which can be calculated using (3).

Leveraging (3) and (4), and assuming variation-free DCs, the MZI transfer matrix in (1) can be updated to take into consideration the impact of FPVs on MZI arms, resulting in undesired optical phase noises:

$$T'_{MZI}(\theta,\phi) = \begin{pmatrix} \frac{\sqrt{2}}{2}e^{i(\theta+\Delta\beta_3 L_3)} & i\frac{\sqrt{2}}{2}e^{i\Delta\beta_4 L_4} \\ i\frac{\sqrt{2}}{2}e^{i(\theta+\Delta\beta_3 L_3)} & \frac{\sqrt{2}}{2}e^{i\Delta\beta_4 L_4} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2}e^{i(\phi+\Delta\beta_1 L_1)} & i\frac{\sqrt{2}}{2}e^{i\Delta\beta_2 L_2} \\ i\frac{\sqrt{2}}{2}e^{i(\phi+\Delta\beta_1 L_1)} & \frac{\sqrt{2}}{2}e^{i\Delta\beta_2 L_2} \end{pmatrix}, \tag{5}$$

where, as shown in Fig. 2, $L_1 = L_2 = L_\phi$ and $L_3 = L_4 = L_\theta$. In this article, we assume $L_\phi = L_\theta \approx 135$ μm, considered as an example based on [14]. Leveraging (5), we can capture the impact of non-uniform variations in the waveguide width, SOI thickness, and slab thickness in any MZI design in coherent SPNNs. Note that FPVs will deviate the splitting ratio, which ideally should be 50:50, in the input and output directional couplers (DC1 and DC2 in Fig. 2), hence affecting the network accuracy [4]. However, this article focuses on the optical phase noise due to FPVs in MZIs and considers variation-free DCs in SPNNs. Also, note that the analyses proposed in this section are independent of the example MZI considered in Fig. 2.

B. Network Level: OIU Performance Under FPVs

Here, we model the impact of FPVs on the performance of an OIU, shown in Fig. 1(b). Note that in this article we do not consider the impact of FPVs in the OAUs and NAUs (see Fig. 1(a)) as they are often implemented either electronically or opto-electronically [2], [15]. Recall from Section II-B that, given a weight matrix $W = U\Sigma V^H$, the matrices U, Σ, and V can be decomposed to θ and ϕ phase values on each MZI in an OIU using Clements decomposition [9]. Under FPVs, the transfer matrix of each MZI in the OIU will deviate in a manner that can be calculated using (5). To analyze the impact of such variations at the network level, we use relative-variation distance (RVD). RVD determines the deviation between an intended matrix and a deviated matrix [4]. We found that there is a strong correlation between network-level RVD and system-level inferencing accuracy in SPNNs (we will discuss this in Section V). RVD can be defined as:

$$RVD\left(W,\overline{W}\right) = \frac{\Sigma_m\Sigma_n\left|W^{m,n} - \overline{W}^{m,n}\right|}{\Sigma_m\Sigma_n\left|W^{m,n}\right|}, \tag{6}$$

where $|.|$ denotes the absolute value of a complex number. W is the nominal weight matrix and $\overline{W}$ is the deviated weight matrix under FPVs related to a fully connected layer. $W^{m,n}$ denotes the element at the m^{th} row and n^{th} column of W. Each MZI in the OIU has a unique impact on each element of the weight matrix. Accordingly, variations related to each MZI in the network have a unique effect on the overall RVD. Higher RVD means the actual weight matrix is more deviated from the intended one, which can be interpreted as observing a lower inferencing accuracy at the system level.

To compute $\overline{W}$ in (6), we first need to analyze the impact of non-uniform FPVs in the waveguide width, SOI thickness, and slab thickness on each individual MZI in an OIU. As the first step, we calculate the total dimension of an OIU based on the length of an individual MZI (l_{MZI}) and the distance between its input and output ports (g_{MZI}). In this work and as an example, we consider $l_{MZI} = 340$ μm and $g_{MZI} = 30$ μm, and based on the phase shifter length in [14] (i.e., $\approx$135 μm). Accordingly and using the Clements design for the OIU [9], the total dimension of the OIU can be calculated. As the second step, we use our in-house FPV wafer map simulator (see Section III-A) to generate a die map that matches the size of OIU. We then place the OIU on the die map and extract FPV information for each individual MZI in the OIU. Note that each MZI itself experiences different variations (non-uniform variations across a single device; a.k.a. intra-device variations), and the FPVs between two different MZIs are also different. All such non-uniformities are considered in our device-level (Section III-A) and network-level analyses. By capturing FPVs for each individual MZI in an OIU,

we can calculate the deviated $\overline{W}$ in (6) based on:

$$\overline{W}_m = \overline{U}_m \times \overline{\Sigma}_m \times \overline{V_m^H}. \tag{7}$$

Here, $\overline{U}_m$ and $\overline{V_m^H}$ are the deviated unitary transfer matrices and $\overline{\Sigma}_m$ is the deviated diagonal matrix under FPVs. They can be calculated based on multiplying MZI transfer matrices under FPVs (the model in (5)), and in a specific order determined by the Clements design [9]. For example, for $\overline{U}_m$ we have:

$$\overline{U}_m = D\left(\prod_{(m,n)\in S} T'_{MZI,m,n}\right), \tag{8}$$

where m and n should be calculated based on the mapping method (e.g., Clements [9]) used to map the weight matrices to cascaded MZI arrays. Also, S is the order of multiplication which again should be determined by the mapping method. Moreover, D is a diagonal matrix with unity magnitude and is not related to the physical placement of MZIs. Similarly, we can calculate $\overline{V_m^H}$ and $\overline{\Sigma}_m$. Although we considered the Clements method for mapping the weights to phase settings of a cascaded MZI array in OIUs, our network-level models in this section can work with any mapping method.

IV. SPNN Design Optimization Under FPVs

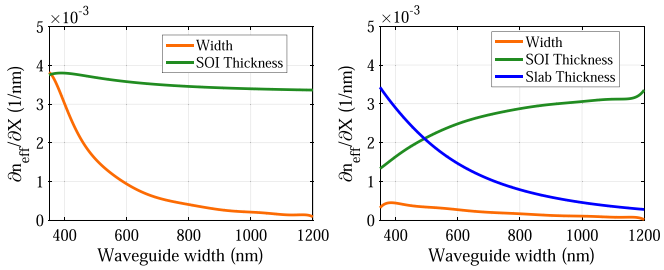
In this section, we explore the design space of MZI devices under different FPVs to optimize their performance in SPNNs. In particular, we focus on minimizing the impact of FPVs on MZI arms that imposes undesired optical phase noises in the device, leading to faulty matrix-vector multiplication. As discussed in Section III, FPVs also deviate the splitting ratio in DCs in an MZI. Nevertheless, the design optimization solution in this section assumes ideal DCs. Note that FPV-tolerant DCs can be designed based on the method proposed in [16].

Considering (4), one can alleviate the impact of FPV-induced optical phase noise in an MZI by implying $|\Delta\beta_1 L_1 - \Delta\beta_2 L_2| \to 0$ and $|\Delta\beta_3 L_3 - \Delta\beta_4 L_4| \to 0$. Accordingly, to obtain a phase-noise-free MZI, we should have:

$$\frac{L_1}{L_2} = \frac{\Delta\beta_2}{\Delta\beta_1} \text{ and } \frac{L_3}{L_4} = \frac{\Delta\beta_4}{\Delta\beta_3}. \tag{9}$$

This indicates that the length ratio between any two opposite arms in the MZI should be inversely proportional to the ratio of the changes in their waveguide propagation constants ($\Delta\beta$), under non-uniform FPVs. In this section and for brevity, we assume $L_1 = L_2 = L_3 = L_4$ and without variations. As a result and based on (9), we should minimize $|\Delta\beta_1 - \Delta\beta_2|$ and $|\Delta\beta_3 - \Delta\beta_4|$ under different FPVs. In other words, we should make sure that the propagation constant changes on the two opposite arms in an MZI are as small as possible (i.e., $\Delta\beta_1 \to 0$, $\Delta\beta_2 \to 0$, $\Delta\beta_3 \to 0$, and $\Delta\beta_4 \to 0$), or the propagation constant changes on the two opposite arms are as close as possible (i.e., $\Delta\beta_1 \to \Delta\beta_2$ and $\Delta\beta_3 \to \Delta\beta_4$).

Considering (3), $\Delta\beta$ is proportional to the rate of changes in the waveguide's effective index under FPVs (i.e., $\frac{\partial n_{eff}}{\partial X}$, where X denotes the design parameter under FPVs: i.e., $X = w$ for width, $X = t$ for SOI thickness, and $X = h$ for slab thickness



(a) Strip waveguide (b) Shallow-etched ridge waveguide

Fig. 3. Rate of changes in waveguide effective index (see the strip and ridge waveguides in Fig. 2) under FPVs $\frac{\partial n_{eff}}{\partial X}$, where X shows the design parameter under FPVs, in (a) a strip and (b) a shallow-etched ridge waveguide, when the waveguide width (w) increases from 350 to 1200 nm. Results are for $t = 220$ nm and $h = 150$ nm (for the ridge waveguide in (b)).

variations. In our prior work [13], we found that as the waveguide width increases, $\frac{\partial n_{eff}}{\partial X}$ decreases, especially under waveguide width variations (i.e., when $X = w$). This is because as the waveguide width increases, a bigger portion of the optical mode is confined in the waveguide core (and the confinement is also stronger), and hence the variations in the waveguide width will create less distortion in the optical mode in the waveguide. Also, note that increasing the waveguide width helps reduce the propagation loss in strip and ridge waveguides [13]. Fig. 3(a) shows the rate of changes in the effective index of a strip waveguide with $t = 220$ nm (see Fig. 2) and when the waveguide width (w) changes from 350 nm to 1200 nm, both considered as an example. As it can be seen, as the waveguide width increases, $\frac{\partial n_{eff}}{\partial w}$ decreases sharply but $\frac{\partial n_{eff}}{\partial t}$ decreases sightly and stays higher than $\frac{\partial n_{eff}}{\partial w}$ under different waveguide widths. While waveguide width can be changed during the design time, the SOI thickness cannot be changed; this parameter is determined by the host SOI wafer.

As it can be seen from Fig. 3(a), both $\frac{\partial n_{eff}}{\partial t}$ and $\frac{\partial n_{eff}}{\partial w}$ are still high in strip waveguides. To address this, we also explore the design of MZIs using a shallow-etched ridge waveguide with a slab thickness (h) of 150 nm, as shown in Fig. 2. We simulated different slab thicknesses from 60 nm to 180 nm (results are not shown in the article), and $h = 150$ nm returned the best results. Such a ridge waveguide is common in the design of grating structures [17]. By adding the slab to a strip waveguide (i.e., making it a ridge waveguide), the optical mode is pulled mostly towards the slab region, hence SOI thickness variations should have less impact on the optical mode. Similar to Fig. 3(a), Fig. 3(b) shows the rate of changes in the effective index of the shallow-etched ridge waveguide with $t = 220$ nm and $h = 150$ nm, and when the waveguide width (w) changes from 350 nm to 1200 nm. Observe that, compared to the strip waveguide, both $\frac{\partial n_{eff}}{\partial w}$ and $\frac{\partial n_{eff}}{\partial t}$ are smaller in the ridge waveguide. The shallow-etched ridge waveguide also suffers from variations in its slab thickness. Nevertheless, as it is shown by Fig. 3(b), $\frac{\partial n_{eff}}{\partial h}$ is much smaller than $\frac{\partial n_{eff}}{\partial t}$ in the shallow-etched ridge waveguide, and it decreases sharply as the waveguide width increases.

Considering the results in Fig. 3, designing MZIs with wider strip and shallow-etched ridge waveguides should help minimize

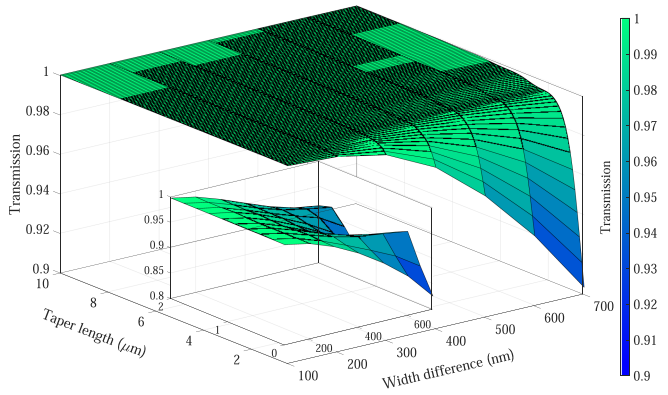


Fig. 4. Minimum taper length required to keep the optical transmission between two waveguides of different widths consistent (i.e., at 1 in the figure) and to avoid mode distortion. The inset zooms in the results for the taper length of 0–2 μm.

the changes in the propagating constant on each MZI arm (see (3) and (9)). Moreover, to increase the waveguide width, an important design consideration is to include waveguide tapers on the MZI arms as shown in Fig. 2. Waveguide tapers are essential to avoid optical mode distortion and higher order mode excitation when moving from the nominal waveguide width (i.e., 470 nm in this paper—see Section V) to a wider waveguide and vice versa [18]. In particular, the taper length should be long enough to avoid any optical transmission and mode distortion. Using Lumerical MODE [19], we simulated the fundamental mode transmission between two waveguides of different widths in Fig. 4. As it can be seen, a waveguide taper length of $\approx$1 μm will be sufficient for every 100 nm width difference between two waveguides of different widths (see Fig. 2). This helps us calculate the area overhead when we optimize MZIs with wider waveguide widths.

Considering different FPVs in the waveguide width, SOI thickness, and slab thickness, modeled based on FPV wafer map models in [13], we consider two scenarios based on which the design of MZIs in an SPNN can be optimized. First, in the *region-based-tolerant MZI design*, we assume a designer may have some *a priori* knowledge of the FPVs. This is a valid assumption as silicon photonics foundries can provide some FPV maps, with different variation data resolutions, to the designers using their fabrication processes. Second, we assume a *worst-case-tolerant MZI design* scenario, where a designer may have very little to no *a priori* knowledge of the FPVs, and hence the MZIs in an SPNN are designed considering the worst-case FPV scenarios (e.g., corner analysis). In such a scenario, we design the worst-case-tolerant MZIs with the largest possible waveguide widths, and equal widths on all the arms, while considering the area overhead in the MZIs. This is discussed further in Section V.

For the region-based-tolerant MZI design, we can define different regions of different sizes (i.e., number of MZIs) in an SPNN, an example of which is shown in Fig. 5. Such regions group the MZIs that are spatially close on the die with one another. We assume that the designer has some *a priori* knowledge of the FPVs for all (and not the individual) MZIs

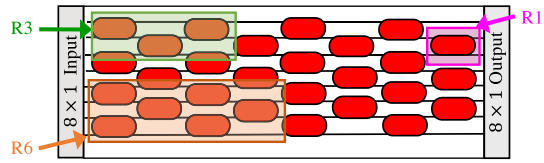


Fig. 5. Different region sizes (R1, R3, and R6) and related MZIs in a single 8 $\times$ 8 OIU unit. R12 (not shown) can be obtained similarly. Each MZI block size is 30 $\times$ 340 μm^2.

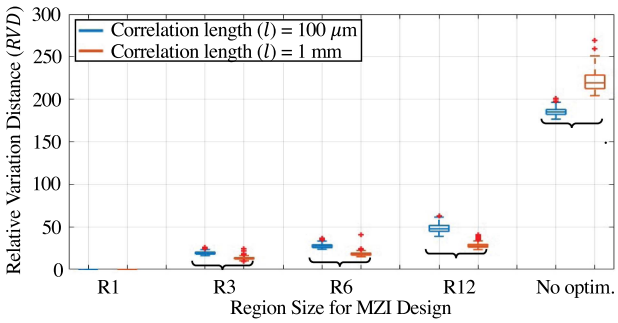
grouped in the same region. As a result, the smaller the region size (e.g., R1 with a single MZI in Fig. 5), the more detailed FPV information is available to the designer and vice versa (e.g., R6 with six MZIs in Fig. 5). Accordingly, we consider the average observed non-uniform variations in a region to design region-based-tolerant MZIs by performing an exhaustive search for the MZI waveguide widths (using results in Fig. 3) while considering the area overhead in the MZIs imposed by adding the tapers. Here, we might have different waveguide widths (between the search range of 350 nm to 1200 nm—see Fig. 3) on each MZI arm.

As we will show in Section V, our region-based-tolerant and worst-case-tolerant MZI designs minimize optical phase noises in MZIs under different FPVs, hence they improve the network accuracy in SPNNs. Nevertheless, it is important to note that our device-level design optimization solutions proposed in this section do not aim at completely eliminating, if at all feasible, the impact of FPVs in SPNNs. That being said, our optimization will reduce the impact of FPVs in SPNNs sufficiently so that the overhead and complexity of dynamic calibration techniques (e.g., [6]) to eliminate the impact of such variations in SPNNs will be significantly reduced.

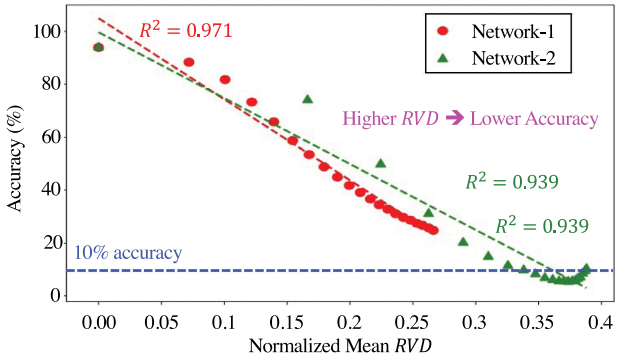
V. Simulation Results and Discussions

FPVs lead to undesired optical phase noises, which, in turn, lead to faulty matrix-vector multiplication in the fully connected layers of an SPNN. The sensitivity of a standalone MZI to FPVs depends on its physical design parameters (see Fig. 3), among which only the MZI waveguide width can be freely altered during the design time. Leveraging realistic and correlated FPV maps developed based on [13] (see Fig. 2-top) and the proposed MZI design optimization in Section IV, we explore and optimize the nominal waveguide widths in the MZIs in SPNN case studies considered in this section, to improve their tolerance under different FPVs.

Prior to evaluating the impact of the proposed MZI optimization on SPNN accuracy under FPVs, let us explore whether such an optimization is independent of the model and the dataset. To examine this, Fig. 6(a) considers 100 randomly generated 16 $\times$ 16 unitary matrices—each of which belongs to a different weight matrix—and presents a box plot of the distribution of RVD's between the nominal and the deviated unitary matrices under FPVs, and for different region sizes. Recall from Section IV that to design the region-based-tolerant MZIs, we take the mean variations affecting a region on the FPV map with 1, 3, 6, or 12 MZIs into consideration (see Fig. 5). All the MZIs in a



(a) RVD for 100 random unitary matrices



(b) Accuracy and RVD under FPVs (different σ_w's and σ_t's)

Fig. 6. (a) RVD for different region sizes and under FPVs with different correlation lengths. (b) High R^2 values denote a strong linear correlation between RVD and accuracy. FPVs are based on the parameters in Table I. We consider the linear correlation as an example to calculate R^2.

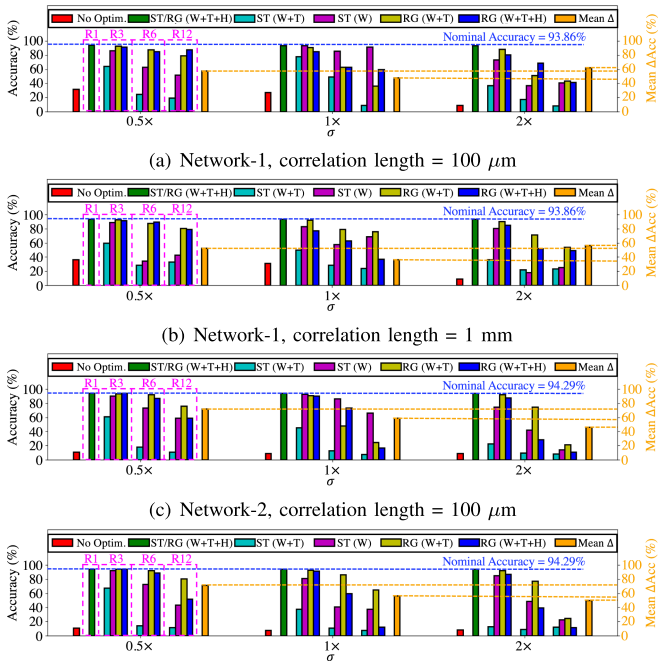
TABLE II
ARCHITECTURES OF THE SPNNS CONSIDERED

Model	Architecture	# PhS
Network-1	FC(16,16)-SP-FC(16,16)-SP-FC(16,10)-LSM	1380
Network-2	FC(64,64)-SP-FC(64,64)-SP-FC(64,10)-LSM	20,580

FC(X,Y): fully connected layer with X inputs and Y outputs, SP: softplus activation, LSM: logsoftmax activation, PhS: phase shifters.

particular region are replaced with the optimized region-based-tolerant MZI, designed using shallow-etched ridge waveguides (see Figs. 2 and 3(b)). As it is shown in Fig. 6(a), in all cases (i.e., R1–R12), the mean RVD is significantly reduced when optimized MZIs are used. In particular, the interquartile ranges (IQRs) in the box plots are consistent among all the unitary matrices in a region: this shows that the proposed optimization is effective independent of the considered unitary matrix. Also, the mean RVD decreases with decreasing the region size (i.e., when a designer has access to more detailed FPV data from the foundry), and it is the highest when MZIs are not optimized.

To explore the impact of our proposed MZI design optimization on SPNN accuracy, we consider a case study of two fully connected SPNNs with different footprint (see Table II) trained on the MNIST dataset. To compress the $28 \times 28 = 784$ dimensional feature vector in the MNIST dataset, we take the shifted fast Fourier transform of each image. The compressed 16-dimensional feature vector for Network-1 is then obtained by considering the values within the 4×4 region at the center of the frequency spectrum. Similarly, for the larger Network-2, we use a 64-dimensional feature vector by considering the 8×8 region at the center of the frequency spectrum. In Fig. 6(b), we show the linear correlation between the accuracy and the mean RVD, averaged over the 6 unitary matrices (two in each of the three OIUs) and normalized over the number of phase shifters in each network. Given such a strong linear correlation, our method should improve the accuracy of all SPNNs under FPVs, irrespective of the nominal phase angles.



(a) Network-1, correlation length = 100 μm

(b) Network-1, correlation length = 1 mm

(c) Network-2, correlation length = 100 μm

(d) Network-2, correlation length = 1 mm

Fig. 7. Accuracy of two SPNNs in Table II under correlated FPVs in width, SOI thickness, and slab thickness before and after the optimization. Here, ST and RG denote strip and shallow-etched ridge waveguide, respectively. The parameters listed inside (.) show the variations considered, with W, T, and H denoting waveguide width, SOI thickness, and slab thickness variations, respectively. Note that results for ST (W+T) and RG (W+T+H) are the same for R1. The second y-axis shows the average difference between No Optim. accuracy (i.e., using the conventional MZI) and the accuracy obtained using RG (W+T+H) for R1, R3, R6, and R12.

By applying realistic and correlated FPV maps—generated using our prior work in [13] and considering parameters in Table I—to Network-1 and Network-2, Fig. 7 shows the inferencing accuracy in each network with conventional MZIs (No Optim.) and optimized region-based-tolerant MZIs, which can have different waveguide widths on each arm. For the conventional MZI, we designed an MZI using strip waveguides with $t = 220$ nm and $w = 470$ nm and variation-free $\approx$10-μm-long DCs with a 200 nm gap, to obtain 50:50 splitting ratio in the MZI. In each plot in Fig. 7, we consider three different standard deviations for the waveguide width, SOI thickness, and slab thickness variations: $0.5\times$, $1\times$, and $2\times$ the expected standard deviations (σ_w, σ_t, σ_h) in Table I. Moreover, the variation maps are generated for two correlation lengths of 100 μm and 1 mm. Considering Fig. 7(a)–(d), with no MZI optimization (i.e., No Optim.; red bars), the SPNN accuracy is always the least (e.g., 7.73% with $1\times\sigma$ in Fig. 7(d)). Considering optimized region-based-tolerant MZI design with strip or shallow-etched ridge waveguides under all the variations and R1 (i.e., when

regions include a single MZI—see Fig. 5), the network accuracy in Fig. 7(a)–(d) is almost the same as the nominal accuracy. This is because each individual MZI has been optimized considering its exact FPV profile. This is for an ideal case when a designer has full access to variation data affecting each MZI in the network, so it may be impractical in most cases (R1 results are shown to indicate the efficiency of our optimization in such rare cases).

Considering more realistic and practical region sizes (R3, R6, and R12) and a region-based-tolerant MZI design using strip waveguides, Fig. 7(a)–(d) show the SPNN accuracy when considering (i) both waveguide width and SOI thickness variations (W+T; light green bars), and (ii) width variations only (W; magenta bars)—when SOI thickness variations are negligible, e.g., through SOI thickness uniformity improvement [20]. Observe that with (i), the optimized MZI can help retrieve some accuracy, which also decreases as the region size increases. But overall, the gain in accuracy in this case is small. This is due to the fact that optimized MZIs designed using strip waveguides are not sufficiently tolerant to thickness variations (see Fig. 3(a)). Considering (ii), the optimized MZI designed using strip waveguides achieves better accuracy improvements (and higher than (i)) in both the networks.

Fig. 7(a)–(d) also show the network accuracies with the optimized region-based-tolerant MZIs designed using shallow-etched ridge waveguides, under the presence of (i) waveguide width and SOI thickness variations (W+T; yellow bars), and (ii) all the variations (W+T+H; dark blue bars). Observe that with both (i) and (ii), the optimized MZIs using shallow-etched ridge waveguides perform much better compared to those using strip waveguides (i.e., ST (W+T)). Comparing (i) and (ii)—when considering slab thickness variations—the network accuracy is (slightly) lower in some cases. Nevertheless, considering the average difference between No Optim. accuracy and the accuracy obtained with (ii) for R1, R3, R6, and R12 (orange bars; second y-axis), using optimized MZIs designed with shallow-etched ridge waveguides under all the variations can significantly improve the network accuracy (e.g., by up to 72% in Network-2 with $0.5\times\sigma$). This shows that the proposed optimization can improve the resilience of SPNNs to FPVs.

Another observation is how the network accuracy in Fig. 7(a)–(d) seem to change across different region sizes and FPV correlation lengths. In fact, when the correlation length in variations is shorter, variations tend to "change more" within a region with a given size, and as the region size increases, they change even more across the region. This imposes a higher error for the region-based-tolerant MZIs designed per region. Therefore, the accuracy results are generally a bit lower in Fig. 7(a) and (c) compared to those in Fig. 7(b) and (d).

To assess SPNN accuracy using the optimized worst-case-tolerant MZI design (see Section IV), we consider, as an example, Network-2 with FPVs of different correlation lengths and standard deviations in Table I. In this experiment, we assume no *a priori* FPV knowledge and using strip waveguides in the design of the optimized worst-case-tolerant MZIs while considering waveguide width variations only. The worst-case-tolerant MZI is optimized by widening all the MZI arms together—i.e., MZI arms all have the same width after optimization—while considering the resulting area overhead due to the required waveguide tapers (see Figs. 2, 3(a), and 4). Results for this experiment (before and after the optimization) are shown in Table III for different area overhead. We observe that even with 1% area overhead, the accuracy improves. However, the improvements are significant only when the area overhead is greater than 8% for both the correlation lengths.

TABLE III
NETWORK-2 ACCURACY WITH WORST-CASE-TOLERANT MZIS DESIGNED USING STRIP WAVEGUIDES UNDER WIDTH VARIATIONS

Area Overhead	Correlation Length	Width (nm)	Arm Length (μm)	Pre-Opt Accuracy	Post-Opt Accuracy
1%	1 mm 100 μm	533	135.63	11.65% 54.27%	20.47% 77.08%
2%	1 mm 100 μm	589	136.19	11.65% 54.27%	45.79% 86.48%
4%	1 mm 100 μm	688	137.18	11.65% 54.27%	83.97% 92.17%
8%	1 mm 100 μm	853	138.83	11.65% 54.27%	92.36% 93.77%
16%	1 mm 100 μm	1111	141.41	11.65% 54.27%	93.97% 94.28%
32%	1 mm 100 μm	1200	142.3	11.65% 54.27%	94.07% 94.29%

VI. CONCLUSION

In this article, we have analyzed the impact of FPVs in the waveguide width, SOI thickness, and slab thickness on coherent SPNNs. In particular, we have modeled undesired optical phase noises due to such variations at the MZI device level, and how such phase noises contribute to the performance degradation, for which we have considered relative variation distance (RVD) at the network level in optical unitary multipliers built using MZIs. Furthermore, we have proposed physical-level design optimization solutions to enhance MZI device tolerance under correlated FPVs in SPNNs. Our simulation results for two SPNN case studies of different sizes and considering realistic FPV maps show that the proposed physical-level optimization can help significantly improve the SPNN inferencing accuracy. In addition, the results in this article indicate the importance of considering variations during the design-phase of SPNNs to facilitate the application of online and dynamic calibration mechanisms in these networks, which are often complex and power- and area-hungry.

REFERENCES

[1] F. P. Sunny, E. Taheri, M. Nikdast, and S. Pasricha, "A survey on silicon photonics for deep learning," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 17, no. 4, pp. 1–57, 2021.
[2] Q. Cheng, J. Kwon, M. Glick, M. Bahadori, L. P. Carloni, and K. Bergman, "Silicon photonics codesign for deep learning," *Proc. IEEE*, vol. 108, no. 8, pp. 1261–1282, Aug. 2020.
[3] E. Cartlidge, "Optical neural networks," *Opt Photon. News*, vol. 31, no. 6, pp. 32–39, 2020.
[4] S. Banerjee, M. Nikdast, and K. Chakrabarty, "Characterizing coherent integrated photonic neural networks under imperfections," *J. Lightw. Technol.*, early access, Aug. 2, 2022, doi: 10.1109/JLT.2022.3193658.
[5] Y. Zhu et al., "Countering variations and thermal effects for accurate optical neural networks," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Des.*, 2020, pp. 1–7.

[6] S. Bandyopadhyay et al., "Hardware error correction for programmable photonics," *Optica*, vol. 8, no. 10, pp. 1247–1255, 2021.
[7] M. Y. et al., "Design of optical neural networks with component imprecisions," *Opt. Exp.*, vol. 27, no. 10, pp. 14009–14029, 2019.
[8] M. Reck et al., "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, pp. 58–61, 1994.
[9] W. R. Clements et al., "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, 2016.
[10] F. Shokraneh et al., "The diamond mesh, A phase-error- and loss-tolerant field-programmable MZI-based optical processor for optical neural networks," *Opt. Exp.*, vol. 28, pp. 23495–23508, 2020.
[11] M. Nikdast, G. Nicolescu, J. Trajkovic, and O. Liboiron-Ladouceur, "Chip-scale silicon photonic interconnects: A formal study on fabrication non-uniformity," *J. Lightw. Technol.*, vol. 34, no. 16, pp. 3682–3695, Aug. 2016.
[12] S. Banerjee et al., "Optimizing coherent integrated photonic neural networks under random uncertainties," in *Proc. IEEE/OSA Opt. Fiber Commun. Conf. Exhib.*, 2021, pp. 1–3.
[13] A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 41, no. 10, pp. 3359–3372, Oct. 2022.
[14] F. Shokraneh, M. S. Nezami, and O. Liboiron-Ladouceur, "Theoretical and experimental analysis of a 4×4 reconfigurable MZI-based linear optical processor," *J. Lightw. Technol.*, vol. 38, no. 6, pp. 1258–1267, Mar. 2020.
[15] M. M. P. Fard et al., "Experimental realization of arbitrary activation functions for optical neural networks," *Opt. Exp.*, vol. 28, no. 8, pp. 12138–12148, 2020.
[16] Z. Lu et al., "Broadband silicon photonic directional coupler using asymmetric-waveguide based phase control," *Opt. Exp.*, vol. 23, no. 3, pp. 3795–3808, 2015.
[17] G. Jiang, R. Chen, Q. Zhou, J. Yang, M. Wang, and X. Jiang, "Slab-modulated sidewall Bragg gratings in silicon-on-insulator ridge waveguides," *IEEE Photon. Technol. Lett.*, vol. 23, no. 1, pp. 6–8, Jan. 2011.
[18] Y. Fu et al., "Efficient adiabatic silicon-on-insulator waveguide taper," *Photon. Res.*, vol. 2, no. 3, pp. A41–A44, 2014.
[19] Lumerical, "Ansys lumerical." Accessed: Apr. 2022. [Online]. Available: https://www.lumerical.com/products/
[20] S. K. Selvaraja et al., "SOI thickness uniformity improvement using corrective etching for silicon nano-photonic device," in *Proc. IEEE Int. Conf. Group IV Photon.*, 2011, pp. 71–73.