Modeling Silicon-Photonic Neural Networks under Uncertainties

Sanmitra Banerjee†, Mahdi Nikdast‡, and Krishnendu Chakrabarty†

†Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
‡Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523, USA

***Abstract*—Silicon-photonic neural networks (SPNNs) offer substantial improvements in computing speed and energy efficiency compared to their digital electronic counterparts. However, the energy efficiency and accuracy of SPNNs are highly impacted by uncertainties that arise from fabrication-process and thermal variations. In this paper, we present the first comprehensive and hierarchical study on the impact of random uncertainties on the classification accuracy of a Mach–Zehnder Interferometer (MZI)-based SPNN. We show that such impact can vary based on both the location and characteristics (e.g., tuned phase angles) of a non-ideal silicon-photonic device. Simulation results show that in an SPNN with two hidden layers and 1374 tunable-thermal-phase shifters, random uncertainties even in mature fabrication processes can lead to a catastrophic 70% accuracy loss.**

I. Introduction

In deep neural networks (DNNs), matrix multiplication is known to be the most time- and energy-intensive operation. Silicon-photonic neural networks (SPNNs) employ photonic components to optimize matrix multiplication with ultra-high speed and ultra-low energy consumption [1]. The linear multipliers are represented using two unitary multipliers and a diagonal matrix, which are obtained using singular value decomposition (SVD). The multipliers and the diagonal matrix can be realized using a network of interconnected Mach–Zehnder interferometers (MZIs) [2]. In the absence of optical crosstalk, the complexity of matrix-vector multiplication can be reduced from $O(N^2)$ to $O(1)$ [1]. However, there exist several roadblocks in the further advancement of SPNNs; these include the optical loss associated with MZI networks [2], [3], additional computation needed for mapping the trained weights to the parameters (i.e., phase angles) in MZI arrays [2], and the finite-encoding precision on phase settings [1].

In this paper, we present the first comprehensive analysis of the impact of uncertainties due to fabrication-process variations (FPVs) and thermal crosstalk in SPNNs. Perturbations in specific MZIs, depending on their position and tuned phase angles, can be catastrophic in nature. Therefore, identifying such components during the design time is necessary for improving the yield. To address this requirement, we develop a framework to identify critical components in SPNNs where random uncertainties lead to severe performance degradation in the network. Significant degradation in SPNN performance (70% loss in inferencing accuracy) is observed considering typical uncertainties—reported in prior work [4]—in the MZIs.

This research was supported in part by the National Science Foundation (NSF) under grant CCF-2006788.

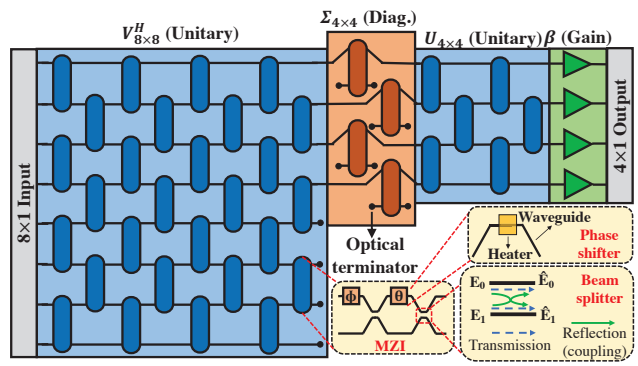


Fig. 1: SPNN linear-layer representation using MZI arrays. An 8×4 linear layer is represented in this example. Bottom: An MZI structure.

II. Background and Motivation

A. Mach–Zehnder Interferometer (MZI)

As shown in Fig. 1, a typical MZI consists of two tunable phase shifters (PhS, ϕ and θ) on the upper arm and two 50:50 beam splitters (BeS). The PhS are used to apply configurable phase shifts and obtain varying degrees of interference between the input optical signals. They can be implemented using thermal microheaters, where the refractive index of the underlying waveguide changes with temperature (i.e., thermo-optic effect), altering the phase of the optical signal traversing the waveguide. Moreover, 2×2 BeS can be designed using directional couplers, where a fraction (defined by transmittance) of the optical signal at an input port is transmitted to an output port, and the remaining (defined by the reflectance) is coupled to the other output port with a phase shift of $\frac{\pi}{2}$. For symmetric 50:50 BeS, both transmittance and reflectance coefficients are $\frac{1}{\sqrt{2}}$. As a result, the transfer matrix for an MZI with two PhS and two 50:50 BeS (see Fig. 1) can be defined as [5]:

$$T_{MZI}(\theta,\phi) = U_{BeS}\cdot U_{PhS}(\theta)\cdot U_{BeS}\cdot U_{PhS}(\phi)$$
$$= \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} = \begin{pmatrix} \frac{e^{i\phi}}{2}(e^{i\theta}-1) & \frac{i}{2}(e^{i\theta}+1) \\ \frac{ie^{i\phi}}{2}(e^{i\theta}+1) & -\frac{1}{2}(e^{i\theta}-1) \end{pmatrix}, \quad (1)$$

where U_{BeS} (U_{PhS}) is the BeS (PhS) transfer matrix.

B. Design of MZI-based SPNNs

Fully connected layers can be represented mathematically as matrix-vector multiplication followed by an activation function. Consider a layer L_i with n_i neurons fully connected to the previous layer L_{i-1} with n_{i-1} neurons. The output vector at L_i is then given by $O_i^{n_i\times1} = f_i(M_i^{n_i\times n_{i-1}} O_{i-1}^{n_{i-1}\times1})$. Note that

f_i and M_i are the non-linear activation function and weight matrix associated with layer L_i, respectively. In SPNNs, the linear multiplication with the weight matrix (i.e., M_i) is often implemented using arrays of configurable MZIs. Using SVD and considering Fig. 1, we have $M_i = U_i \Sigma_i V_i^H$, where U_i and V_i are unitary matrices with dimensions $n_i \times n_i$ and $n_{i-1} \times n_{i-1}$, respectively. Moreover, V_i^H denotes the Hermitian transpose of V_i and Σ_i is a diagonal matrix consisting of the eigenvalues of M_i.

Given a weight matrix $M_i = U_i \Sigma_i V_i^H$, we use the Clements design [2] to represent the unitary matrices U_i and V_i^H. The diagonal matrix Σ_i can be represented using a similar MZI array where one input and one output of each MZI are terminated (see Fig. 1). A global optical amplification is necessary on each output to represent arbitrary diagonal matrices [6]. This scaling factor is realized using layer β, as shown in Fig. 1.

C. Related Work on Component Imprecision in SPNNs

Deviations in the phase angles in PhS and the splitting ratios in BeS—due to inevitable FPVs and thermal crosstalk—have a severe impact on MZI performance in SPNNs [7]. The use of thermal actuators to compensate for phase errors leads to induced mutual thermal crosstalk between neighboring waveguides [8]. A method to counter the impact of both FPVs and thermal effects using a modified cost function during training and post-fabrication hardware calibration is presented in [9]. However, this method only focuses on uncertainties in the phase angles, ignoring the considerable impact of inevitable errors in BeS. Moreover, the required hardware calibration necessitates the tuning of each MZI in the network, and this step becomes increasingly complex as the network scales up. The modified training method also results in accuracy loss.

Here, we model the impact of random and non-uniform uncertainties in both phase angles and beam-splitting ratios in MZIs in SPNNs. We also show that the impact of uncertainties depends both on the position and parameter values of the affected MZIs. Therefore, some random variations in some MZIs can be more critical than others. Our entire analysis can be performed prior to fabrication and after software training.

III. Uncertainties in SPNNs: A Hierarchical Study

In this section, we systematically analyze the impact of uncertainties on the performance of SPNNs in a hierarchical fashion at the component-level (PhS and BeS), device-level (MZIs), layer-level (MZI array), and system-level (SPNN).

A. Component-Level: Phase Shifters and Beam Splitters

The temperature-dependent phase change in a thermo-optic PhS is given by $\Delta\phi = \left(\frac{2\pi l}{\lambda_0}\right) \cdot \left(\frac{dn}{dT}\right) \cdot \Delta T$, where l is the length of the phase shifter and λ_0 is the optical wavelength [10]. Also, $\frac{dn}{dT} \approx 1.8 \times 10^{-4} K^{-1}$ is the thermo-optic coefficient of silicon at $\lambda_0 = 1550$ nm and temperature $T = 300\ K$ [11], and ΔT is the temperature change. During *in-situ* training of SPNNs, the phase angles at PhS are applied using thermal actuators (i.e., microheaters). Mutual thermal crosstalk among neighboring actuated waveguides, which are placed in proximity in SPNNs (see Fig. 1), affects the efficiency of the tuning and bias-control mechanism, imposing phase-angle errors. Furthermore, FPVs can change l (see $\Delta\phi$), hence impacting the efficiency of PhS. Due to random perturbations in the phase angles (θ and ϕ) in (1), T_{MZI} will deviate from its intended form, resulting in faulty matrix multiplication and a reduction in SPNN inferencing accuracy.

Considering the classical, lossless 2×2 beam-splitter schematic shown in Fig. 1 the electric fields at the output $\tilde{E}_{0/1}$ can be attributed to the transmitted electric-field component E_0 and the reflected electric-field component E_1 based on [5]:

$$\begin{pmatrix} \tilde{E}_0 \\ \tilde{E}_1 \end{pmatrix} = \begin{pmatrix} r_{00} & it_{10} \\ it_{01} & r_{11} \end{pmatrix} \begin{pmatrix} E_0 \\ E_1 \end{pmatrix}. \tag{2}$$

Here, r and t represent the reflectance and transmittance associated with each path, respectively. Note that $r_{00}^2 + t_{01}^2 = 1$ and $r_{11}^2 + t_{10}^2 = 1$. For symmetric BeS, $r_{00} = r_{11} = r$ and $t_{01} = t_{10} = t$. Additionally, for ideal 50:50 BeS, $r = t = \frac{1}{\sqrt{2}}$. However, under random FPVs, r and t will deviate from $\frac{1}{\sqrt{2}}$; this results in unbalanced and imperfect BeS [12], [13]. Unlike PhS, BeS are passive devices and once fabricated, we cannot actively change their r and t values during SPNN training.

Prior studies have shown an error of ~0.21 radian in the tuned phase angles in PhS for mature fabrication processes [4]. This corresponds to $\frac{0.21}{2\pi} \times 100 \approx 3.34\%$ of the range of phase angles. Taking this into consideration, we perturb θ and ϕ using a Gaussian distribution with mean (μ) set to their nominal tuned values (obtained from training) and multiple values of standard deviation in the range $0.005 \cdot 2\pi \leq \sigma \leq 0.15 \cdot 2\pi$. While a deviation of 1–2% is typically expected in the r and t parameters in BeS [4], we vary them using a similar distribution as PhS—Gaussian with $\mu = \frac{1}{\sqrt{2}}$ and $0.005 \cdot \frac{1}{\sqrt{2}} \leq \sigma \leq 0.15 \cdot \frac{1}{\sqrt{2}}$—for a fair comparison of their impact on accuracy. In the rest of the paper, we use σ_{PhS} to refer to $\frac{\sigma}{2\pi}$ for PhS, and σ_{BeS} to refer to $\sqrt{2}\sigma$ for BeS.

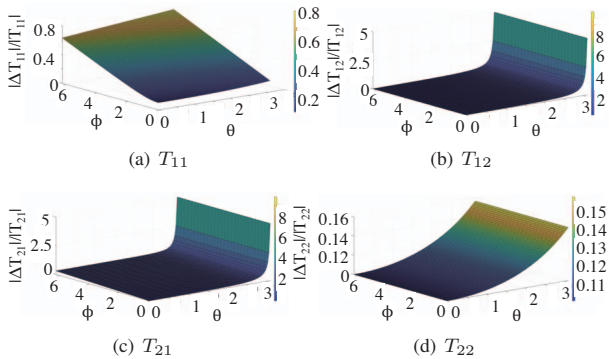
B. Device-Level: MZIs

Variations in θ ($\Delta\theta$) and ϕ ($\Delta\phi$) phase angles in PhS can result in deviations in the MZI transfer matrix, T_{MZI}, defined in (1). Such deviations can be defined as:

$$\begin{aligned} \Delta T_{MZI}(\theta,\phi) &= \frac{\partial T_{MZI}(\theta,\phi)}{\partial\theta}\Delta\theta + \frac{\partial T_{MZI}(\theta,\phi)}{\partial\phi}\Delta\phi \\ &= \begin{pmatrix} \frac{ie^{i(\phi+\theta)}}{2} & -\frac{e^{i\theta}}{2} \\ -\frac{e^{i(\phi+\theta)}}{2} & -\frac{ie^{i\theta}}{2} \end{pmatrix} \Delta\theta + \begin{pmatrix} \frac{ie^{i\phi}}{2}(e^{i\theta}-1) & 0 \\ -\frac{e^{i\phi}}{2}(e^{i\theta}+1) & 0 \end{pmatrix} \Delta\phi. \end{aligned} \tag{3}$$

Let the relative changes in θ and ϕ be $K_\theta = \frac{\Delta\theta}{\theta}$ and $K_\phi = \frac{\Delta\phi}{\phi}$, respectively. We assume $K_\theta = K_\phi = K$ as the two PhS, corresponding to θ and ϕ, are in proximity (see Fig. 1). Note that this assumption is made to simplify the analyses only in this subsection. In all subsequent analyses, independent variations are considered in θ and ϕ. Thus, from (3), we have:

$$\Delta T_{MZI}(\theta,\phi) = K \begin{pmatrix} (\theta+\phi)\frac{ie^{i(\theta+\phi)}}{2} - \phi\frac{ie^{i\phi}}{2} & -\theta\frac{e^{i\theta}}{2} \\ -(\theta+\phi)\frac{e^{i(\theta+\phi)}}{2} - \phi\frac{e^{i\phi}}{2} & -\theta\frac{ie^{i\theta}}{2} \end{pmatrix}. \tag{4}$$



(a) T_{11} (b) T_{12} (c) T_{21} (d) T_{22}

Fig. 2: Magnitude of variation in the absolute value of elements in T_{MZI} (see (1)) relative to the modulus of their nominal values.

Using (1) and (4), Fig. 2 shows the magnitude of deviation for each of the four elements in T_{MZI} relative to the modulus of their nominal values for different values of θ and ϕ with $K = 0.05$. We find that the relative deviation increases monotonically as θ and ϕ increase. This indicates that MZIs with higher values of tuned phase angles are more susceptible to uncertainties.

The proposed T_{MZI} model in (1) assumes ideal 50:50 BeS with $r_{00} = r_{11} = t_{01} = t_{10} = \frac{1}{\sqrt{2}}$. However, under uncertainties in BeS, this model changes to:

$$T_{MZI}(\theta,\phi) = \begin{pmatrix} rr'e^{i(\theta+\phi)} - tt'e^{i\phi} & ir'te^{i\theta} + it'r \\ it're^{i(\theta+\phi)} + itr'e^{i\phi} & -tt'e^{i\theta} + rr' \end{pmatrix}. \quad (5)$$

Here, r (t) and r' (t') are the reflectances (transmittances) for the first and the second beam splitter, respectively (see Fig. 1).

C. Layer-Level: MZI Array

Under uncertainties, T_{MZI} deviates, and consequently, the matrix represented by the array can vary from the intended unitary matrix. We use the relative-variation distance (RVD) as a figure-of-merit to quantify the difference between the intended unitary matrix ($\tilde{U}$) and the deviated unitary matrix (U). This is given by $RVD(U,\tilde{U}) = \frac{\sum_m \sum_n |U_{m,n} - \tilde{U}_{m,n}|}{|\tilde{U}_{m,n}|}$.

Different elements of the unitary transfer matrix are affected by different subsets of MZIs in the array. Therefore, variations in each MZI will have a unique impact on the overall RVD defined above. This is indeed the case as is shown in Fig. 3. We consider four randomly generated 5×5 unitary matrices with random perturbations in the PhS and BeS. For each matrix, we introduce variations in one MZI at a time. For each MZI, we perform 1000 Monte Carlo iterations and calculate the average RVD. In each iteration, the MZI parameters (θ, ϕ, r, r', t, t') corresponding to the faulty MZI are chosen from a Gaussian distribution with $\sigma_{PhS} = \sigma_{BeS} = 0.05$. From Fig. 3 we observe that there is a significant variation in the average RVD corresponding to different MZIs representing the same unitary matrix. Note also that the distribution of average RVD across the MZIs differs across the four unitary matrices. Thus, it is clear that the impact of uncertainties in the MZI array on the accuracy of the unitary multiplier varies from case to case.

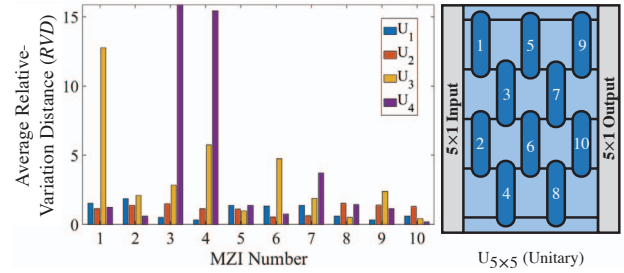


Fig. 3: Average RVD (left) for four random 5×5 unitary matrices with one MZI under variations at a time. Right: An MZI array (including the MZI numbers) to represent any 5×5 unitary matrix (see Fig. 1).

D. System-Level: SPNN

Variations in the MZI parameters lead to faulty matrix multiplications in the linear layers, imposing classification accuracy loss in SPNNs. To show the severe impact of such variations in SPNNs, we present a case study of an SPNN handling the MNIST hand-written digit classification task [14].

To convert the 28×28 = 784 dimensional real-valued images in the MNIST dataset to complex-valued vectors, we consider the shifted fast Fourier transform of each image; this results in a 784-dimensional complex-valued vector for each image. To compress the feature vector, we consider the values within 4×4 region at the center of the frequency spectrum. Compared to the baseline accuracy of 94.12% with the 28×28 feature vector, the 4×4 case results in only 6.77% accuracy loss.

In our SPNN architecture, fully connected feedforward networks with two hidden layers of 16-complex valued neurons are implemented using the Clements design [2]. Each linear layer is followed by the nonlinear Softplus function applied to the modulus of the complex numbers. To model intensity measurement, a modulus squared nonlinearity is applied after the output layer. This is followed by a final LogSoftMax layer to obtain a probability distribution. We use a cross-entropy loss function during training [15].

We realize the three weight matrices corresponding to the neurons in the input and the two hidden layers in our SPNN using MZI arrays. Based on our network architecture, the dimensions of the weight matrices are 16×16 (input layer), 16×16 (first hidden layer), and 16×10 (second hidden layer). To analyze the impact of random uncertainties in the MZI arrays on the SPNN, we perform the following experiments:

- EXP_1 (global uncertainties): We select a σ_{PhS} and σ_{BeS} and for each selected value, perform 1000 Monte Carlo iterations. For each iteration, we calculate the inferencing accuracy using the 10000 test images in the MNIST dataset. The use of 1000 Monte Carlo iterations is formally justified based on the fact that with a 95% confidence interval, the maximum margin of error in the mean of the inferencing accuracy is 6.27%, which is within the acceptable range [16]. Note that EXP_1 is performed with uncertainties inserted only in PhS, only in BeS, and in both where $\sigma_{PhS} = \sigma_{BeS}$.
- EXP_2 (global uncertainties with zonal perturbations): To find the impact of localized uncertainties on the SPNN accuracy, we divide the SPNN into different zones, each

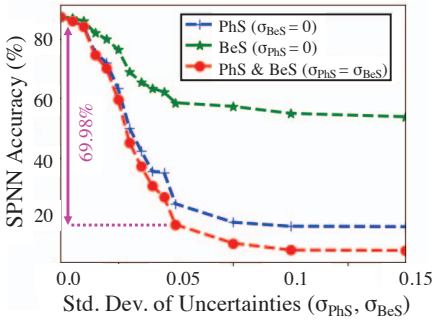


Fig. 4: Impact of random uncertainties in the SPNN components (PhS and BeS) on the SPNN inferencing accuracy (EXP_1).

consisting of four MZIs arranged in a 2×2 grid. We insert random perturbations with $\sigma_{PhS} = \sigma_{BeS} = 0.1$ in a selected zone while the remaining zones have uncertainties with $\sigma_{PhS} = \sigma_{BeS} = 0.05$. For each selected zone, we again consider 1000 Monte Carlo iterations (similar to EXP_1) and calculate the reduction in the mean inferencing accuracy from the nominal case.

Fig. 4 shows the simulation results for EXP_1 when uncertainties are inserted in (i) only PhS, (ii) only BeS, and (iii) both PhS and BeS. For all these cases, the accuracy declines steeply as σ increases before it saturates around $\sigma = 0.075$ where the accuracy drops below 10% (accuracy associated with a random guess). Also, we can see that uncertainties in PhS have a higher impact on accuracy compared to those in BeS.

The three linear layers in our SPNN can be represented by six unitary multipliers. The impact of zonal perturbations in these unitary multipliers on the classification accuracy (experiment EXP_2) is presented as heatmaps in Fig. 5. Figs. 5(a)–(b) correspond to the U and V^H matrices of the first linear layer while Figs. 5(c)–(d) and Figs. 5(e)–(f) correspond to the second and third linear layers, respectively. Note that for all these cases, the diagonal matrix Σ is assumed to be error-free with the singular values arranged in random order. Each box in the heatmaps corresponds to a zone with the height (width) of the layer increasing vertically (horizontally). The value (color) in each box signifies the accuracy loss when a zonal perturbation is applied to the corresponding zone. From experiment EXP_1 (Fig. 4), we know that the reduction in SPNN accuracy under a global uncertainty of $\sigma_{PhS} = \sigma_{BeS} = 0.05$ is 69.98%. Fig. 5 shows that even under zonal perturbations, the accuracy loss hovers around 69.98%. However, in some zones, the zonal perturbations result in a decreased accuracy loss (e.g., the zone in row 2 column 5 in Fig. 5(a)), whereas in others they exacerbate the impact of global uncertainties (e.g., the zone in row 3 column 0 in Fig. 5(f)). Moreover, note that the low- and high-impact zones are arranged randomly in each unitary multiplier. This shows that the impact of localized uncertainties in MZIs can differ significantly and some MZIs are more critical than others (see also Fig. 3).

IV. Conclusion

We have modeled the impact of random uncertainties in SPNNs that arise due to fabrication-process variations and thermal crosstalk. Simulation results from our hierarchical approach show that even minor uncertainties in SPNN building blocks have a significant impact on the inferencing accuracy and reliability in SPNNs. Such impact depends on both the tuned parameter values and the position of affected components. The proposed modeling framework can be used to identify and compensate for critical components in SPNNs during design.

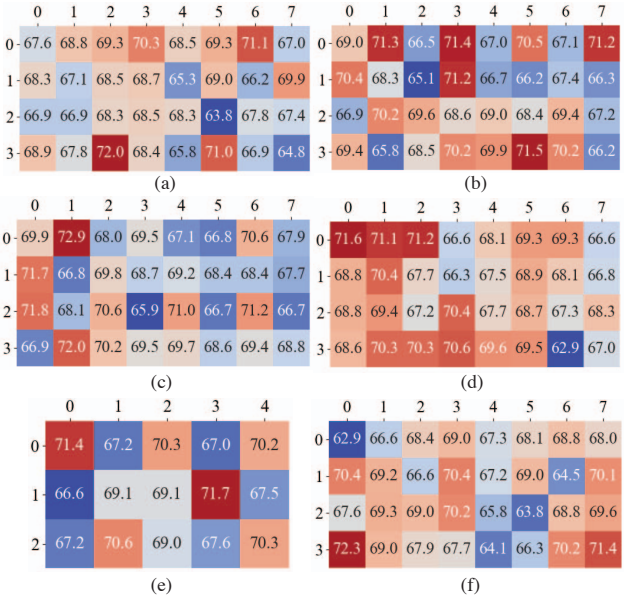


Fig. 5: Accuracy loss (%) due to zonal perturbations in linear layers (EXP_2): (a) U_{L0}, (b) V^H_{L0}, (c) U_{L1}, (d) V^H_{L1}, (e) U_{L2}, and (f) V^H_{L2}.

References

[1] Q. Cheng *et al.*, "Silicon photonics codesign for deep learning," *Proceedings of the IEEE*, vol. 108, no. 8, pp. 1261–1282, 2020.
[2] W. R. Clements *et al.*, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, 2016.
[3] M. Reck *et al.*, "Experimental realization of any discrete unitary operator," *Physical Review Letters*, vol. 73, no. 1, p. 58, 1994.
[4] F. Flamini *et al.*, "Benchmarking integrated linear-optical architectures for quantum information processing," *Scientific Reports*, 2017.
[5] M. Y. S. Fang *et al.*, "Design of optical neural networks with component imprecisions," *Optics Express*, pp. 14 009–14 029, 2019.
[6] M. J. Connelly, *Semiconductor Optical Amplifiers*. Springer, 2007.
[7] Z. Lu *et al.*, "Performance prediction for silicon photonics integrated circuits with layout-dependent correlated manufacturing variability," *Optics Express*, vol. 25, no. 9, pp. 9712–9733, 2017.
[8] M. Milanizadeh *et al.*, "Canceling thermal cross-talk effects in photonic integrated circuits," *IEEE JLT*, vol. 37, no. 4, pp. 1325–1332, 2019.
[9] Y. Zhu *et al.*, "Countering variations and thermal effects for accurate optical neural networks," *IEEE ICCAD*, pp. 1–7, 2020.
[10] M. Jacques *et al.*, "Optimization of thermo-optic phase-shifter design and mitigation of thermal crosstalk on the SOI platform," *Optics Express*, vol. 27, no. 8, pp. 10 456–10 471, 2019.
[11] D. F. Walls *et al.*, *Quantum Optics*. Springer, 2007.
[12] Y. C. Liu *et al.*, "Compensation of non-ideal beam splitter polarization distortion effect in michelson interferometer," *Optics Communications*, vol. 361, pp. 153–161, 2016.
[13] M. Nikdast *et al.*, "Chip-scale silicon photonic interconnects: A formal study on fabrication non-uniformity," *IEEE JLT*, vol. 34, no. 16, pp. 3682–3695, 2016.
[14] Y. LeCun. (1998) The MNIST database of handwritten digits. [Online]. Available: http://yann.lecun.com/exdb/mnist/.
[15] T. M. Cover *et al.*, *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2006.
[16] (2008) An online survey on statistical significance. [Online]. Available: http://www.surveystar.com/startips/oct2008.pdf.