Phase-change Programmable Photonics for Optical Computing and Signal Processing

Changming Wu

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Mo Li, Chair

Arka Majumdar

Sajjad Moazeni

Program Authorized to Offer Degree:

Electrical and Computer Engineering

University of Washington

Abstract

Phase-change Programmable Photonics for Optical Computing and Signal Processing

Changming Wu

Chair of the Supervisory Committee:
Mo Li
Department of Electrical and Computer Engineering

The programmability in integrated photonic systems fosters advancements across diverse technologies, from data centers to optical neural networks and quantum information processing. Phase-change materials (PCMs) can offer an ideal solution thanks to their reversible switching, large index contrast, and non-volatile behavior, enabling programmability with no static power consumption. In this thesis, I will mainly introduce several phase-change photonic devices that can contribute to various photonic applications such as optical computing, signal processing, and optical communications.

First, we demonstrate a multimode photonic computing core consisting of an array of programable mode converters based on on-waveguide metasurfaces made of phase-change materials. We demonstrate a prototypical optical convolutional neural network that can perform image processing and recognition tasks with high accuracy. With a broad operation bandwidth and

a compact device footprint, the demonstrated multimode photonic core is promising for large-scale photonic neural networks with ultrahigh computation throughputs.

Then we demonstrate a photonic generative network as a part of a generative adversarial network (GAN) that can generate a handwritten number in experiments. We realize an optical random number generator derived from the amplified spontaneous emission noise, apply noise-aware training by injecting additional noise, and demonstrate the network's resilience to hardware non-idealities. Our results suggest the resilience and potential of more complex photonic generative networks based on large-scale, realistic photonic hardware.

Finally, we report direct-write and rewritable photonic circuits based on a low-loss phase change material (PCM) thin film, in which complete end-to-end functional photonic circuits can be created by direct laser writing in one step without additional fabrication processes. The direct-write phase-change photonic circuit affords exceptional flexibility, allowing any part of the circuit to be erased and rewritten, facilitating rapid design modification and reprogramming. We demonstrate the versatility of this technique with various photonic circuits for diverse applications, including an optical interconnect fabric for reconfigurable networking, a photonic crossbar array as a tensor core for optical computing, and a tunable optical filter for optical signal processing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

As time flies by, it has come to the end of my doctorate journey. Throughout the entirety of my Ph.D. journey, I faced numerous setbacks and challenges. At the same time, I also experienced the most gratifying, enriching, and joyful moments of my life so far in Prof. Mo Li's group. And I am filled with deep gratitude for the invaluable support and guidance I have received over more than five years.

Foremost, I express my sincere gratitude to my advisor, Prof. Mo Li, for his unwavering guidance and support. Over the past six years, I engaged in numerous insightful discussions with Mo, which has been instrumental in shaping my research taste and interests. Moreover, I have been deeply impressed by his unwavering passion for research, and I am confident that this influence will continue to shape my future career. Moreover, I am thankful for Mo's genuine concern for every student's goals, coupled with his willingness to provide support whenever needed. I am truly grateful for the endless support that Mo has provided throughout both my research and career endeavors.

Secondly, I extend my heartfelt appreciation to my family, my parents, and my wife, for being companions throughout my entire Ph.D. journey. The patience in listening to my post-work complaints, especially during times when I met failures in both research and life, has been an invaluable source of support. Beyond the realm of research, they are my greatest pillars of strength, and for that, I am truly thankful.

Additionally, I want to express my appreciation to Dr. Huan Li, Dr. He Li, Dr. Ruoming Peng, Dr. Seokhyeong Lee, Dr. Han Zhao, Dr. Che Chen, and Dr. Qiyu Liu—former colleagues in our group—each of whom demonstrated exceptional talent and diligence. In particular, Ruoming and Li He gave me significant support. I cherished the time spent working alongside them, and their contributions have significantly enriched my research experience. Also, I thank my current groupmates: Dr. Bingzhao Li, Adina Ripin, Qixuan Lin, I-Tung Chen, Haoqin Deng, and Shucheng Fang. Their diverse expertise in various facets of photonics has provided valuable insights, and they generously share their research experiences, contributing to a comprehensive understanding of current photonics research.

DEDICATION

This thesis is dedicated to my family

For their unwavering support and endless love that made this journey possible.

Chapter 1. Introduction

Modern photonic systems heavily rely on programmability [1], a critical feature essential for various cutting-edge technologies such as next-generation data centers [2–4], optical computing [5–10], quantum information processing [11–15], and light detection and ranging [16–21]. Achieving programmability traditionally involves employing modulation methods like the thermo-optic (TO) effect [22–24], free-carrier dispersion [25–28], electro-optic (EO) effect [29–31], and mechanical tuning [32,33]. While these physical effects are adept at modulating light, their suitability for programmability remains limited.

For applications requiring infrequent switching, the individual power and latency of switching events become less critical compared to zero static energy retention in programmed states. In such a context, phase-change materials (PCMs), emerge as an ideal solution due to their reversible, bistable phase transition [34], multilevel operation [35,36], and significant refractive index contrast [37]. These materials facilitate rapid programming (~ns to μs) and moderate switching energy (fJ to nJ) [38–40].

This dissertation aims to delve into our endeavors to expand versatile phase-change programmable photonics designs for modern photonic applications, push this technology closer to modern applications for photonic platforms, and outline future research directions in PCM-based photonics. Rather than providing an extensive literature review, this thesis will primarily focus on our original contributions.

1.1 Optical phase-change material

Phase-change materials (PCM) are types of materials characterized by an unconventional combination of properties. In general, PCM materials have two or more stable structural phases, for example, the crystalline phase and the amorphous phase. The amorphous and crystalline phases possess significant differences in electrical and (or) optical properties. At the same time, these alloys can be rapidly and reversely switched between these two phases through external thermal excitation [37,41]. As shown in Figure 1.1, when the PCM is heated over the melting temperature T_m, and suddenly cooling down with the cooling rate larger than 10^9 K/s, the PCM is quenched into the amorphous phases. When the PCM is heated over the glass-transition temperature T_g but below T_m, the PCM is annealed to the crystalline phase. Thus, such characteristics are ideally suitable for applications in programmable data storage [42].

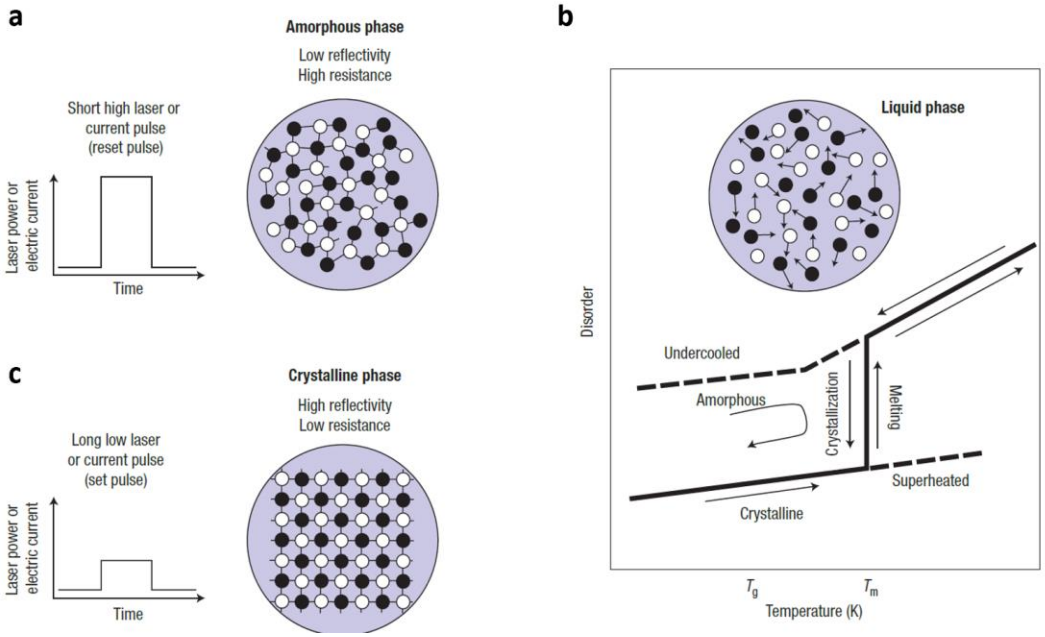

Figure 1.1 Rewriteable optical data storage using phase-change materials. **a.** A short high optical/electrical pulse is used to quench the PCM into the amorphous phase. **b.** The thermal process shows the quenching and annealing process to switch the PCM between phases. A large contrast in the optical properties is shown between amorphous and crystalline phases. **c.** A long low optical/ electrical pulse is used to anneal the PCM into the crystalline phase. The figure is extracted from ref [42].

1.2 Programmable phase-change photonic device

Phase-change materials have emerged as promising candidates for integrated photonics due to their unique optical properties and rapid phase transitions [41]. The commonly used materials, like GST [43,44], and Sb_2Se_3 [45–47], exhibit a reversible change in their optical properties between amorphous and crystalline states when triggered by external stimuli like heat, light, or electrical pulses.

When PCM is integrated with the photonic component, the difference in optical properties in the two phases will introduce the different performance of the photonic device. Figure 1.2 plots the TE_0 mode profile (E_x component) of a 20 nm-thick fully amorphous and crystalline phase GST deposited on the silicon nitride ridge waveguide. Since the GST has a much higher refractive index and extinction coefficient compared to silicon nitride, the electromagnetic field prefers to accumulate in the material with a larger refractive index, a very thin layer of GST as well as its phase can greatly affect the mode profile and optical properties give different n_{eff} and κ_{eff}. This difference in n_{eff} and κ_{eff} will be key in designing the phase-change photonic devices.

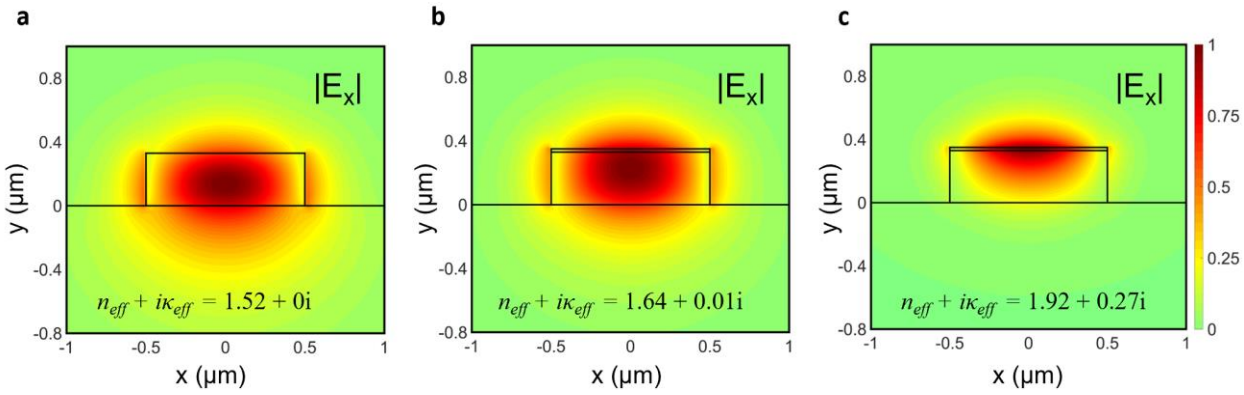

Figure 1.2 **a.** the TE_0 mode profile ($|E_x|$) for a Si_3N_4 ridge waveguide on the SiO_2 substrate. The dimension of the waveguide crosssection is 1 μm (width) × 330 nm (thickness). **b** and **c.** The TE_0 mode profile ($|E_x|$) with the same waveguide geometry but with a 20 nm-thick aGST (**b**) and

cGST (**c**) deposited on top of the waveguide. The waveguide is propagating along the z-direction and the optical wavelength is 1550 nm.

In integrated photonics, phase-change materials are utilized in various ways:

1. Optical switching: PCMs can be integrated into photonic circuits as active components for optical switches [44,48–50]. By exploiting their ability to transition between states quickly, they enable the development of reconfigurable optical switches. This facilitates the redirection of light signals within a photonic chip, enabling efficient routing and modulation of optical data (see Figure 1.3).

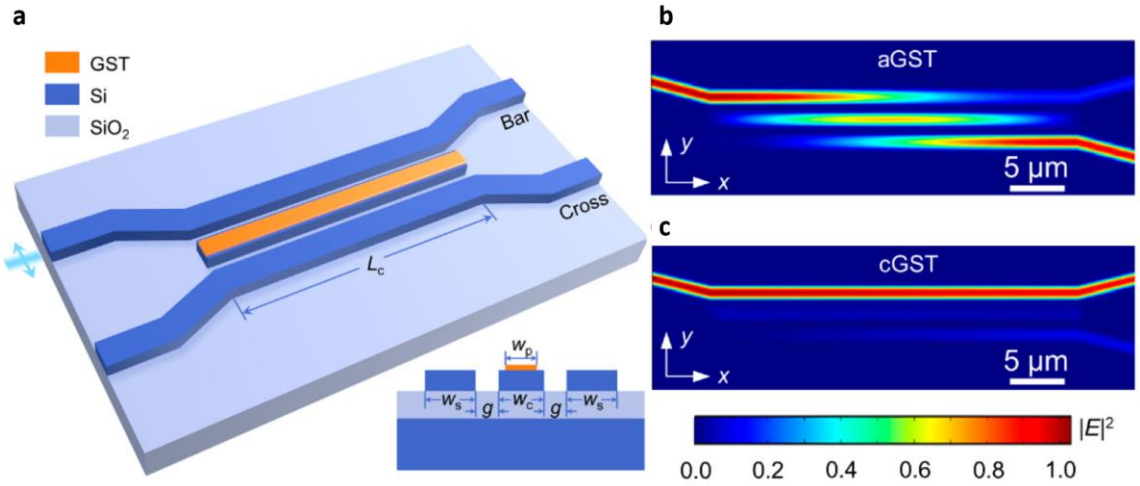

Figure 1.3 Design of a 2 × 2 DC switch. **a.** Schematic of the switch. **b** and **c.** Normalized optical field intensity distribution in the device for aGST (**a**) and cGST (**c**) simulated by the 3D eigenmode expansion method (Lumerical) at 1550 nm. The figure is extracted from ref [50].

2. Reconfigurable photonic circuits: The reversible phase transitions of PCMs allow for the dynamic alteration of optical properties (see Figure 1.4). This feature is harnessed to create reconfigurable photonic circuits [51], enabling adaptive functionalities in data processing, optical computing [52], and telecommunications systems.

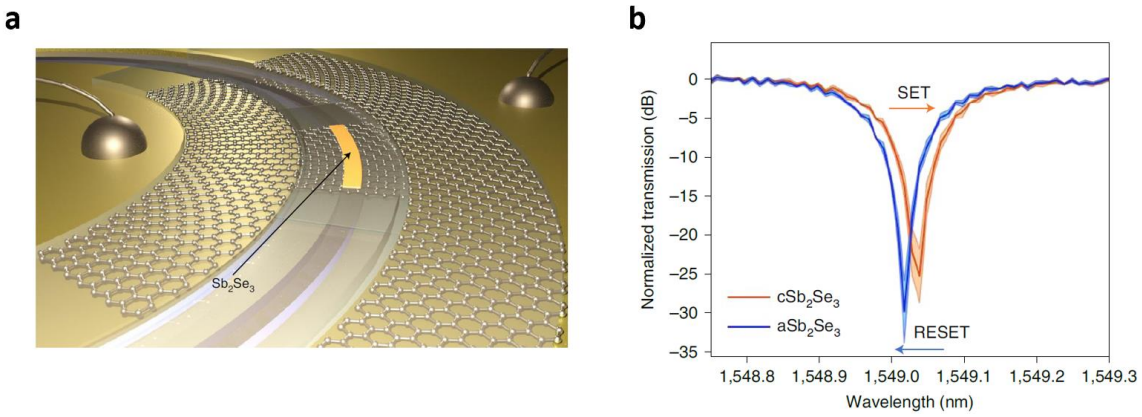

Figure 1.4 **a.** Schematic of the graphene–Sb_2Se_3 phase shifter in a micro-ring. **b.** Reversible switching of Sb_2Se_3 using a graphene heater on micro-rings. Three consecutive cycles are plotted; the shaded areas indicate the standard deviation of the cycles and the solid lines indicate the average. The figure is extracted from ref [53].

3. Data storage: Phase-change materials have long been used in optical data storage devices like CDs and DVDs due to their ability to store and retrieve data by switching between amorphous and crystalline states. Integrated photonics explores the potential of utilizing these materials for on-chip data storage and processing, enabling compact and high-speed photonic memory devices, even in the development of photonic neural networks. As shown in Figure 1.5, the multi-bits of the information can be stored within the transmission level of the phase-change integrated photonics devices and is read out through the optical approach.

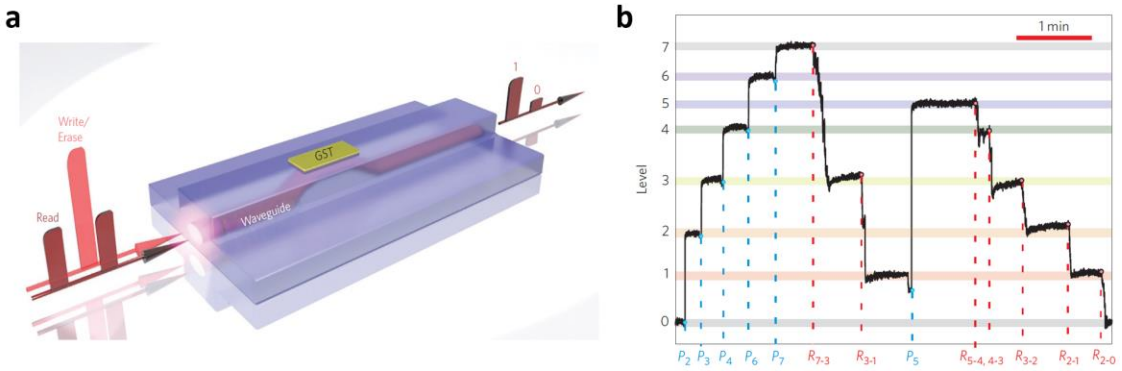

Figure 1.5 **a.** Information is stored in the phase state of the GST section on top of the nanophotonic waveguide. Both reading and writing of the memory can be performed with ultrashort optical pulses because the guided light interacts with the GST via its evanescent field. In the readout, data are encoded in the amount of optical transmission through (along) the waveguide because the two crystallographic states of the GST exhibit a high contrast in optical absorption. **b.** The device is operated with eight levels. The figure is extracted from ref [35].

Integrating phase-change materials into photonic circuits opens up new possibilities for creating versatile, compact, and energy-efficient devices with enhanced functionalities, paving the way for advancements in high-speed data communication, computing, sensing, and beyond within the realm of integrated photonics.

1.3 The scope of this thesis

This dissertation consists of 5 chapters.

In Chapter 2, we demonstrate a multimode photonic computing core consisting of an array of programable mode converters based on on-waveguide metasurfaces made of phase-change materials. The programmable converters utilize the refractive index change of the phase-change material $Ge_2Sb_2Te_5$ during phase transition to control the waveguide spatial modes with a very high precision of up to 64 levels in modal contrast. We further demonstrate a prototypical optical convolutional neural network that can perform image processing and recognition tasks with high accuracy. We show the demonstrated multimode photonic core is promising for large-scale photonic neural networks with ultrahigh computation throughputs.

In Chapter 3, we follow the phase-change material-based photonic computing platform and we focus on developing strategies to mitigate and, if possible, harness noises in photonic computing systems. We demonstrate a photonic generative network as a part of a generative adversarial network (GAN). We demonstrate that the GAN can generate a handwritten number ("7") in experiments and a full ten digits in simulation. We realize an optical random number generator derived from the amplified spontaneous emission noise, apply noise-aware training by injecting additional noise, and demonstrate the network's resilience to hardware non-idealities.

In Chapter 4, we show direct-write and rewritable photonic circuits based on a low-loss phase change material thin film, in which complete end-to-end functional photonic circuits can be created by direct laser writing in one step without additional fabrication processes. The direct-write phase-change photonic circuit affords exceptional flexibility, allowing any part of the circuit to be erased and rewritten, facilitating rapid design modification and reprogramming. We show the versatility of this technique with various photonic circuits for optical computing and a tunable optical filter for optical signal processing.

In the last chapter (Chapter 5), we summarize the dissertation work and provide insights for future applications of the techniques we have shown in this thesis.

Chapter 2. Programmable Phase-change Metasurface on Waveguides for Multimode Photonic Convolutional Neural Network

The widening disparity between the rate of energy efficiency improvements in current digital electronics and the accelerating demand for computation, particularly driven by emerging applications such as machine learning and artificial intelligence, has rekindled interest in optical computing [54,55]. Integrated photonics stands out as a promising hardware platform, enabling the realization of extensive optical networks on a single chip and affords an enormous bandwidth density that is unreachable for electronics [56,57]. To use integrated photonics for optical computing, programmable photonic components, and nonlinear elements are indispensable building blocks [6,58]. Phase-change materials recently emerged as an ideal material system to realize optical programmability. The optical properties of PCMs change dramatically during the phase transition, which can be electrically or optically controlled. Harnessing this has allowed for embodiments of programmable optical switches, couplers, lenses, and metamaterials to be demonstrated [59–61]. Notably, the phase changes in the chalcogenide family of Ge-Sb-Te alloys exhibit nonvolatility, obviating the need for a continuous power supply to maintain the programmed state or stored information. This characteristic offers a significant advantage in power consumption over electro-optic or thermo-optic methods. Photonic devices incorporating these nonvolatile PCMs can thereby serve as optical memories and conduct in-memory computing [7,62], simply by assessing the transmission of optical input data through the programmed device.

In this chapter, we report a programmable waveguide mode converter based on a phase-gradient metasurface made of GST [63]. This phase-change metasurface mode converter (PMMC) utilizes GST's large refractive index change during its phase transition to control the conversion of the waveguide's two spatial modes (TE_0 and TE_1 modes). The PMMC can be programmed to control the waveguide mode contrast precisely at 64 distinguishable levels, which is used to represent the weight parameters with 6-bit precision in matrix-vector multiplication (MVM) computation. We build a 2×2 array of PMMCs and implement them as programmable kernels to realize a multimode optical convolutional neural network (OCNN). By performing image processing tasks such as edge detection and pattern recognition, we demonstrate the OCNN's viability and potential in large-scale optical computing.

2.1 Working principle and PMMC design

The design of the PMMC is based on the principle of a phase-gradient metasurface [64] but replacing noble metals with phase-change materials. Figure 2.1a shows a 3D schematic of the design, which consists of a linear array of GST nano-antennae directly integrated on a silicon nitride waveguide. Each GST nano-antenna scatters the waveguide mode and causes a phase shift Φ, which depends on its geometry (e.g., width), as well as the refractive index of its material. A linear array of such nano-antennae with tapering widths thus produces a spatial gradient of the scattering phases $d\Phi/dx$, which is equivalent to a wavevector k_g. If the phase-gradient metasurface is designed such that k_g matches the wavevector difference between two spatial modes of the waveguide: k_{mode1} - k_{mode2}, it satisfies the phase-matching condition and facilitates the conversion between the two modes. Such phase-gradient metasurface for waveguide mode conversion realized with noble metals or dielectrics materials thus lacked tunability. Here, we use GST, which has a large change in its optical properties when a phase transition happens. When the GST is in the amorphous phase, its refractive index n is ~ 4.7. In contrast, when it is turned to the crystalline phase, n increases to ~7.5 with a drastic change of 2.8 over the whole measured spectral range from 1540 nm to 1580 nm [65]. This change will significantly modify the scattered phase of each GST nano-antenna to modify the metasurface's function.

To build a well-defined phase gradient $d\Phi/dx$, the phase response of a single GST nano-antenna is simulated first. The inset of Figure 2.1b shows the cross-section of the geometry, in which a single GST nano-antenna is placed on a 1.8 μm wide, 330 nm thick Si_3N_4 multimode waveguide, with a 400 nm offset from the central axis of the waveguide. The waveguide and the GST nano-antenna are conformally covered with a layer of Al_2O_3 layer as a protection layer. A fundamental waveguide TE_0 mode is launched into the waveguide. The field distribution right after the TE_0 mode passes through the nano-antenna is recorded. To precisely determine the scattered phase, we conduct an additional simulation using a device with a similar geometry, but without a GST nano-antenna, serving as a reference. This field obtained from the reference is then subtracted from the initial simulation to isolate the contribution of the nano-antenna. This methodology allows us to ascertain both the phase and amplitude information of the scattered field generated by the nano-antenna. Since cGST has a higher refractive index near 1550 nm, the field scattered by

cGST nano-antenna is much stronger compared to the field scattered by aGST nano-antenna (please refer to Figure 2.1c and Figure 2.1d).

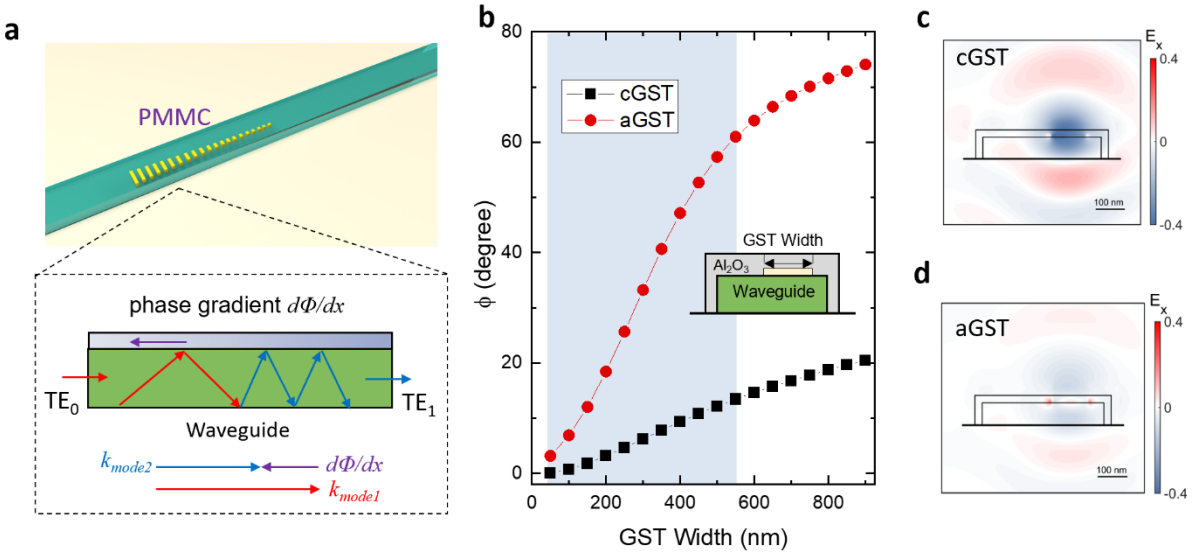

Figure 2.1. The working principle of the PMMC. **a**. 3D illustration of the devices. Inset: Schematic of the phase-matching condition of the phase-gradient devices. **b.** The phase of the scattered mode as a function of the GST nano-antenna width for cGST and aGST phases. The shaded region indicates the range of antenna widths that are used in the phase gradient metasurface. Inset: a cross-sectional view of the structure. **c and d**. Finite element simulation of the scattered electric field by one nano-antenna when the GST is in cGST (**c**) and aGST (**d**) phases, respectively, showing the distinctive difference.

In the simulation, we sweep the width of the nano-antenna as well as the phase of the GST (both aGST and cGST). Figure 2.1b plots the simulated phase of the scattered fields inside the waveguide by a single nano-antenna of 30-nm-thick GST as a function of its width and for aGST and cGST phases. The phase response also shows a much stronger dependence on nano-antenna width when the GST is in the crystalline phase compared to when the GST is in the amorphous phase. Thus, by controlling the geometry of the GST nano-antennae and the interval between adjacent ones in the array, a well-defined phase gradient $d\Phi/dx$ is established when GST is in its crystalline phase. Moreover, considering both the phase and the amplitude response together, a phase-gradient metasurface mode converter designed for the cGST nano-antenna won't be effective when the GST is in the aGST phase, other than causing a small perturbation to the mode.

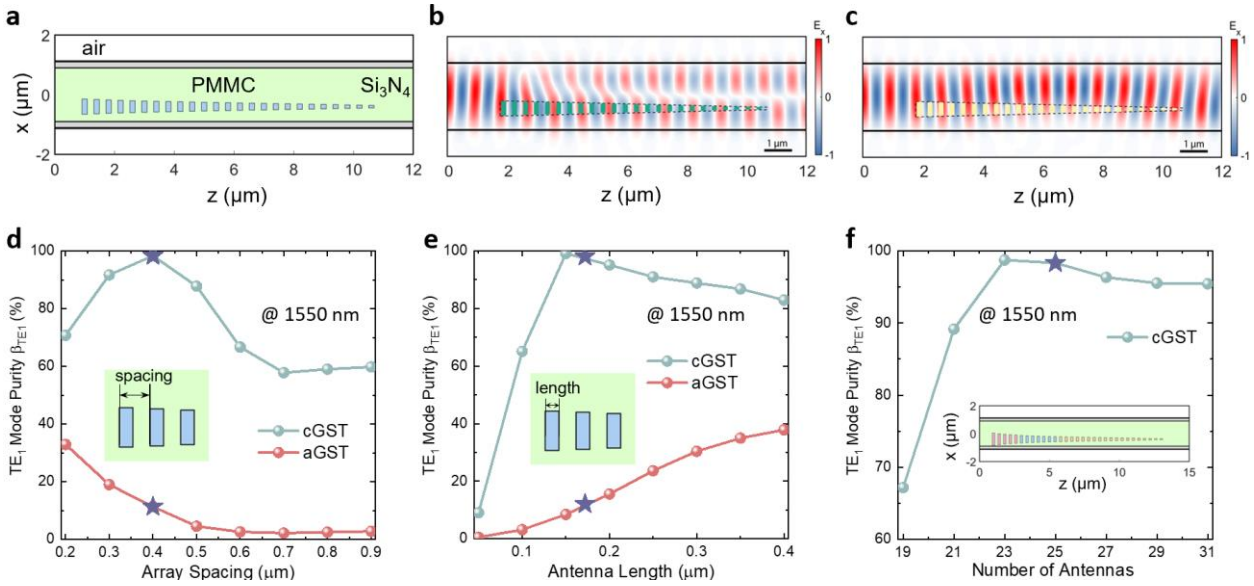

Figure 2.2. Design of the phase-gradient metasurface mode converter. **a**. Top view (x-z cross-sectional plane) of the tapered antennae array. The green, blue, grey, and white areas denote the Si_3N_4 waveguide, GST array, Al_2O_3 capping layer, and ambient air, respectively. **b and c.** FDTD simulation results show effective mode conversion from the TE_0 mode to the TE_1 mode when the GST is in (**b**) crystalline phase, but only a small perturbation when the GST is in (**c**) amorphous phase. **d and e.** Calculated TE_1 mode purity for both cGST and aGST phases, as a function of the antenna length (**d**) and antenna spacing (**e**), The star marks the parameters of the fabricated device. The inset denotes the definition of the parameters. **f.** The TE_1 mode purity as a function of the number of antennae while keeping the phase gradient as a constant. The inset is the geometry of 19 (yellow), 26 (yellow+blue), and 31 (yellow+blue+red) nano-antennae, respectively. The star symbol marks the number of nano-antennas (25) in the fabricated device.

As shown in Figure 2.2a, the metasurface we designed consists of an array of 25 nano-antennae with tapering widths from 510 nm to 84 nm (shaded region in Figure 2.1b). The waveguide supports two transverse-electric modes: the fundamental TE_0 mode and the first-order TE_1 mode. When antennae are in the cGST phase, they induce a uniform change in phase $d\Phi$ of 2.5° for every displacement dx of 400 nm, satisfying the generalized phase-matching condition:

$$k_0\left(n_{\mathrm{TE0}} - n_{\mathrm{TE1}}\right) = N \cdot d\Phi / dx \qquad (2.1)$$

where k_0 is the free space wavevector, n_{TE0} and n_{TE1} are the effective index of the TE_0 and TE_1 modes, respectively, and N is the number of antennae. The cGST metasurface can efficiently convert the TE_0 mode to the TE_1 mode, as confirmed by the simulation result in Figure 2.2b. When the GST is transitioned to the aGST phase, as shown in Fig. 1c, the phase gradient $d\Phi/dx$ is much

reduced, rendering it insufficient to meet the phase-matching condition. Consequently, mode conversion between TE_0 and TE_1 modes does not occur, as evident in Figure 2.2c.

One important parameter to quantify a mode converter's performance is the mode purity in the multimode waveguide, defined as $\beta_{\mathrm{TE0(TE1)}} = \frac{P_{\mathrm{TE0(TE1)}}}{\left(P_{\mathrm{TE0}} + P_{\mathrm{TE1}}\right)}$, where P_{TE0} (P_{TE1}) is the power in the TE_0 (TE_1) mode. As plotted in Figure 2.2d to f, we further conduct several rounds of optimization processes to enhance the TE_1 mode purity of the PMMC. Optimized parameters encompass the thickness of the A_2O_3 encapsulation layer, the lengths of the GST nano-antennae, the spacing dx between adjacent antennae, and the antennae number N. The optimized parameters are listed in Table 2.1.

Table 2.1. Optimized Parameters of PMMC

Central wavelength (nm)	1550
Dimensions of the waveguide cross-section (μm × μm)	1.8 × 0.33
Al_2O_3 layer thickness (nm)	218
Numbers of antennas	25
Antenna length/thickness/spacing (nm)	172/ 30/ 400
Phase incremental (degrees)	2.5
Antenna offset from waveguide central axis (nm)	400
Antenna lengths (nm)	510, 480, 447, 425, 398, 380, 358, 341, 323, 308, 290, 278, 261, 248, 233, 220, 205, 192, 177, 164, 149, 135, 119, 105, 84

2.2 Simulation on the performance of the PMMC

One of the key advantages of employing the GST metasurface lies in its controllability between a fully amorphous and a fully crystalline phase using optical pulses. This characteristic imparts a multilevel mode conversion behavior for the PMMC, facilitating the requisite programmability for reconfigurable photonics and optical computing. Between the full phase transitions, the phase composition of the GST in the metasurface can be continuously tuned by partial phase transition so that the PMMC can be continuously programmed to multiple intermediate levels of mod purity values.

To simulate this multilevel mode conversion of the PMMC, we begin by estimating the effective refractive index n and extinction coefficient κ for GST in intermediate phases (partially crystallized and partially amorphous). As shown in Figure 2.3a, the refractive index n, as well as the extinction coefficient κ of the aGST and cGST are measured using a spectroscopic ellipsometer. The permittivity of GST is then calculated using the effective-medium theory under the effective permittivity approximation:

$$\frac{\varepsilon_{\text{eff}}(p)-1}{\varepsilon_{\text{eff}}(p)+2} = p\times\frac{\varepsilon_c-1}{\varepsilon_c+2}+(1-p)\times\frac{\varepsilon_a-1}{\varepsilon_a+2} \tag{2.2}$$

where ε_{eff} is the effective permittivity of the GST in the intermediate phase, ε_c, and ε_a are the complex permittivities measured using ellipsometry spectroscopy for cGST and aGST phases respectively, and can be obtained from $\sqrt{\varepsilon} = n + i\kappa$, p is the percentage of crystallization such that p=100% corresponds to the fully cGST phase while 0% corresponds to the fully aGST phase. Following the equation, the change of n and κ with the percentage of crystallization p of the GST is calculated (Figure 2.3b).

Based on the optical properties of the GST, assuming continuous and uniform tuning of the GST metasurface phase, Figure 2.4 presents 2D plots of the normalized transmission, TE_1 mode purity, and mode contrast at the output when the TE_0 mode is input to the PMMC, respectively, while varying both wavelength and crystallization percentage. The mode contrast, denoted as Γ, is defined as the difference between TE_0 mode purity (β_{TE0}) and TE_1 mode purity (β_{TE1}), $\Gamma = \beta_{\text{TE0}} - \beta_{\text{TE1}}$, and its theoretical range falls within (-1, 1). The transmission spectra at

different crystalline percentages are normalized against the transmission spectrum at the amorphous phase.

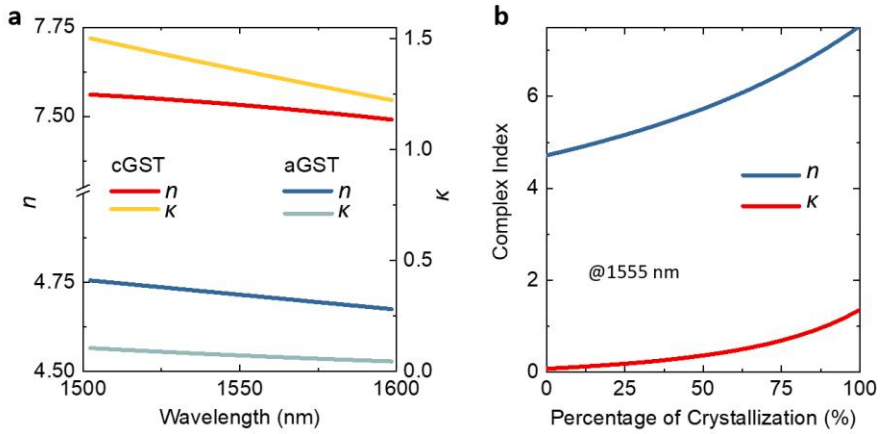

Figure 2.3.Optical properties of the GST. **a**. Refractive index *n* and extinction coefficient κ of the crystalline and amorphous phases of GST measured with ellipsometry. **b.** The change of *n* and κ with the percentage of crystallization *p*.

As illustrated in the simulation, the PMMC demonstrates robust performance across a broadband range from 1500 nm to 1600 nm. As the crystallization percentage increases from 0% to 100%, the normalized transmission is gradually attenuated to ~ 30% due to the cGST's absorption (Figure 2.4a). Despite the lower transmission, in the cGST phase, the TE_1 mode purity exceeds 95% (Figure 2.4b), signifying the desired function that the majority of the transmitted power is carried by the TE_1 mode, corresponding to a mode contrast of -0.9 (Figure 2.4c). Conversely, in the aGST phase, the TE_0 mode purity reaches as high as 87%, indicating that the majority of the transmitted power is borne by the TE_0 mode, and corresponds to a Γ of +0.74. Figure 2.4d to e provides detailed cross-sectional line cuts of the 2D transmission and mode purity plots at a fixed wavelength of 1555 nm. As shown in Figure 2.4f, we assume the crystallization percentage of all antennae uniformly drops from 100% to 0% in 10 steps, with a 10% decrease for each step, the mode contrast undergoes stepwise variations from -1.0 to +0.75, clearly exhibits a multilevel mode conversion behavior.

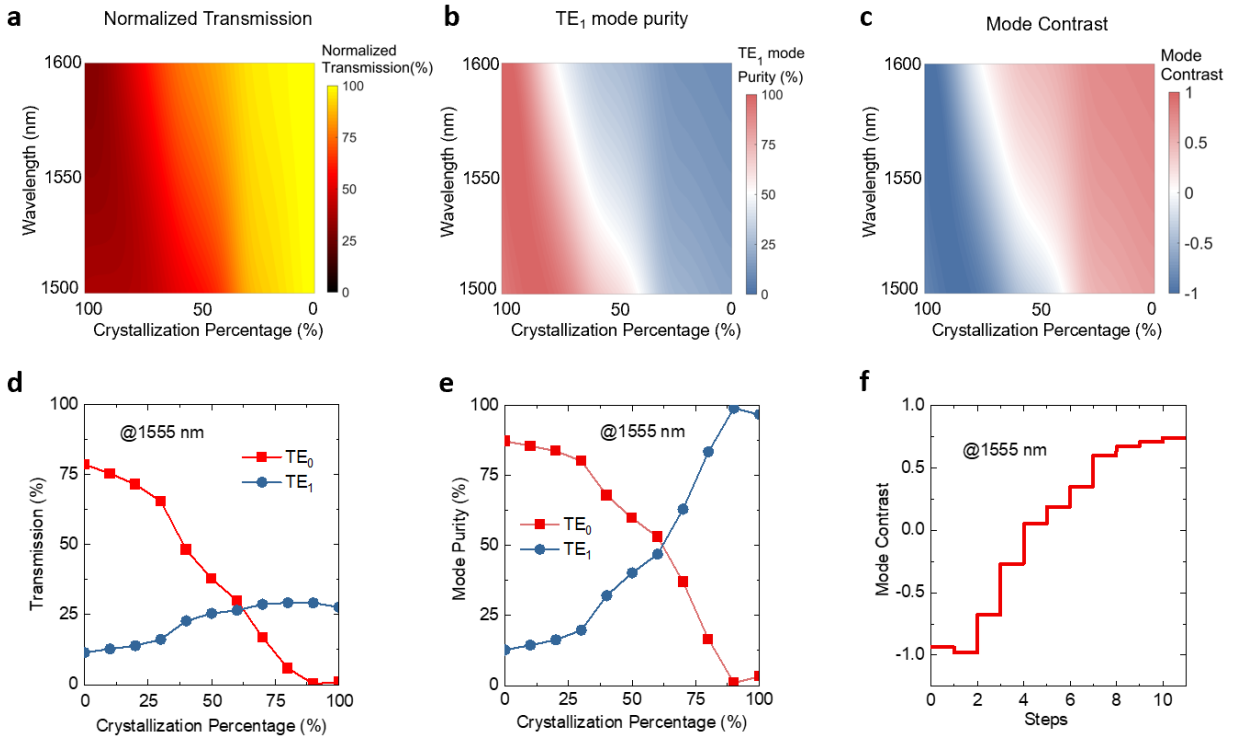

Figure 2.4 Simulating the performance of the PMMC. **a-c.** 2D plots of the normalized transmission (**a**), the TE_1 mode purity (**b**), and the mode contrast (**c**), when both the optical wavelength and the crystallization percentage are scanned. **d-e.** The cross-sectional plot of the transmission (**d**) and the mode purity (**e**) as a function of the crystallization percentage (0% corresponds to cGST and 100% corresponds to aGST). **f.** Simulation of the mode contrast varies when the GST's phase is changed from cGST to aGST step by step. Step 0 corresponds to the fully crystalline phase. In each step, the crystallization percentage drops by 10%. The crystallization percentage drops to 0 (fully amorphous phase) at step 10.

2.3 PMMC fabrication and characterization

To fabricate the PMMC, we deposited a layer of 30 nm thick GST with a layer of 10 nm thick SiO_2 film on top by sputtering on the silicon nitride on the insulator wafer (330 nm thick stoichiometric silicon nitride film deposited by low-pressure chemical vapor deposition (LPCVD) on an oxidized silicon wafer. The metasurface was then patterned with an electron beam lithography system (EBL) using resist ZEP 520A, and etched with an inductively coupled plasma etching (ICP) system using fluorine-based chemistry. Next, the photonic structures are patterned and etched using the beam-lithography (EBL) and dry etching processes. Afterward, the GST nano-antenna array was conformally covered with a 218 nm thick Al_2O_3 layer deposited with the atomic layer deposition (ALD) method followed by a standard lift-off process to complete the fabrication. After fabrication, the substrate was baked at 180 ℃ on a hotplate for 10 minutes to convert the GST into the fully cGST phase.

The deviation between the geometry of the designed and the fabricated nano-antennas is characterized by the SEM images taken before the Al_2O_3 encapsulation process and are listed in Table 2.2. The fabricated width error of each nano-antenna. The fabrication process is optimized to make the nano-antennas in the intermediate part of the array as accurate as possible, with <5% deviation from design. The overall errors in the widths of fabricated nano-antennas are controlled within ~15% except for the shortest nano-antennas. Despite the fabrication errors, the devices have excellent programmable mode conversion performances, demonstrating the robustness of the design against fabrication errors.

To assess the TE_0 and TE_1 mode components at the output following the PMMC, we design an asymmetric directional coupler acting as a mode selector. This component functions by selectively coupling only the TE_1 mode component in the multimode waveguide to the TE_0 mode component in the single-mode waveguide while keeping the TE_0 mode in the multimode waveguide unaffected. For experimental characterization of the mode selector's performance, we fabricated a pair of mode selectors connected with the multi-mode waveguide. The front one serves for inputting the modes, while the back one selects the mode components. As depicted in Figure 2.5, when the TE_1 mode is input, it propagates to the coupling waveguide; conversely, upon input of the TE_0 mode, it remains in the bus waveguide. The transmission efficiency for both modes surpasses 90%.

Table 2.2. The fabricated width error of each nano-antenna

#	Design (μm)	Fabricated (μm)	Error (%)	#	Design (μm)	Fabricated (μm)	Error (%)
1	0.510	0.576	12.94	2	0.480	0.525	9.38
3	0.447	0.505	12.98	4	0.425	0.484	13.88
5	0.398	0.433	8.79	6	0.380	0.428	12.63
7	0.358	0.398	11.17	8	0.341	0.358	4.99
9	0.323	0.353	9.29	10	0.308	0.336	9.09
11	0.290	0.291	0.34	12	0.278	0.288	3.60
13	0.261	0.273	4.60	14	0.248	0.252	1.61
15	0.233	0.241	3.43	16	0.220	0.218	0.91
17	0.205	0.211	2.93	18	0.192	0.202	5.21
19	0.177	0.167	5.65	20	0.164	0.158	3.66
21	0.149	0.135	9.40	22	0.135	0.118	12.59
23	0.119	0.102	14.29	24	0.105	0.066	37.14
25	0.084	0.055	34.52				

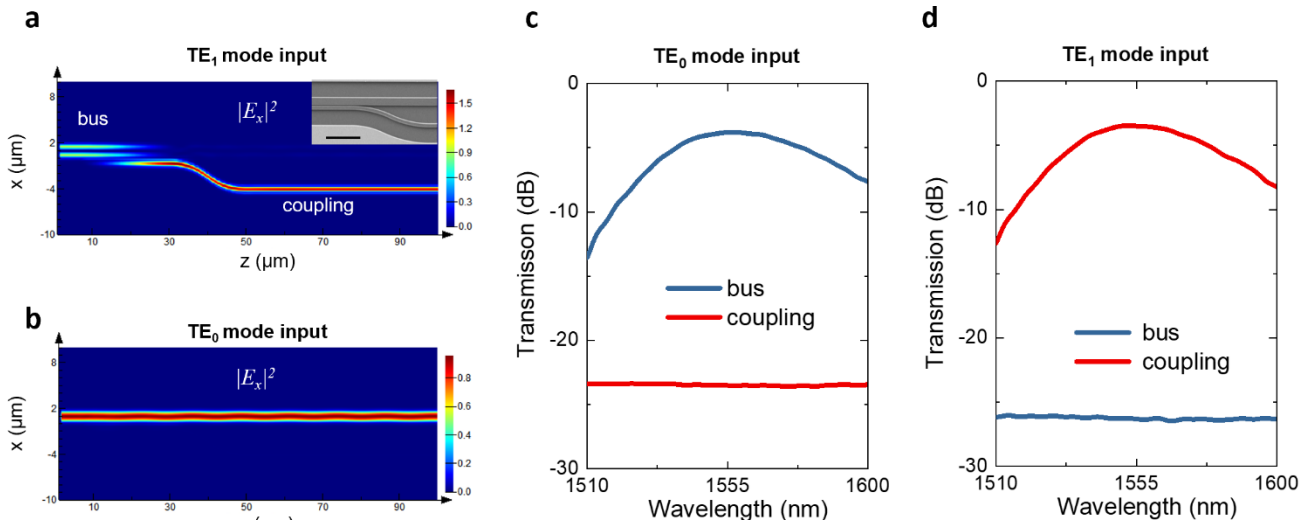

Figure 2.5 **a** and **b**. Simulation on how the mode propagates when passing through the mode selector. If the input mode is TE_1 mode, it will gradually couple to the coupling waveguide (**a**). If the input mode is the TE_0 mode, it will stay in the bus waveguide (**b**). **c** and **d**. Experimental characterization of the mode selector transmission efficiency in each port.

2.4 Operation of the PMMC with high precision

The complete device includes multimode waveguides, mode selectors, and grating couplers. Figure 2.6a depicts the measurement and control scheme. To program the PMMC, we use optical pulses to control the phase of the GST film for simplicity. When operating the PMMC, an optical signal is input in the TE_0 mode to the PMMC and converted to TE_1 mode with a proportion controlled by the state of the GST metasurface. At the output of the PMMC, the TE_1 component is separated by the mode selector and coupled out at the second port while the TE_0 component remains in and outputs from the multimode waveguide. The output powers of both modes are then measured to determine their respective transmission coefficients. Figure 2.6d shows the transmission spectrum of the PMMC when the metasurface is set to be either in fully aGST or cGST phases. The insertion losses of the input and output fibers and grating couplers have been accounted for by calibration measurements. In the aGST phase, the device is in the on-state for the TE_0 mode with a high transmission Ton over a broad wavelength range (1540 to 1580 nm). The lowest insertion loss is 0.9 dB at 1575 nm wavelength. A small portion (<-10 dB) of the TE_1 mode is generated due to the asymmetric perturbation induced by the metasurface even though the aGST phase has a low refractive index.

The situation changes dramatically when the metasurface is transitioned to the cGST phase and converts the $_{TE0}$ mode to the TE_1 mode effectively. In this off-state for the TE_0 mode, its transmission Toff is lower than -15 dB over the entire measured bandwidth. The corresponding switching extinction ratio, defined as $\Delta T / T^{\text{off}} = \left(T^{\text{on}} - T^{\text{off}}\right) / T^{\text{off}}$, is ~16 dB or 4000%, which is more than a 10-fold improvement compared to previously reported switch devices using GST [35,40,49]. This large switching ratio stems from the phase engineering approach to effectively use GST's large refractive index change during its phase transition, as opposed to only using the absorption coefficient change, to facilitate scattering into a different mode that is filtered. The total area of the GST in the metasurface is only 1.3 μm^2, significantly smaller than that in prior devices, and thus in principle, our device consumes less energy to switch. As expected from energy conservation, the TE_1 mode is switched in the opposite way to the TE_0 mode. From aGST to cGST phase, the TE_1 transmission increases from ~-10 dB to ~-6.5 dB, with the insertion loss due to cGST's absorption. Another important parameter to quantify a mode converter's performance is the mode purity which is defined in Section 2.1. The PMMC shows very high

performance in controlling mode purity. As shown in Figure 2.6e, when switching the GST from aGST to cGST phase, the PMMC efficiently converts TE_0 mode to TE_1 mode, changing the mode purity from β_{TE0}>80% to β_{TE1}>85% over a broad bandwidth, showing an excellent agreement with the numerical simulation results.

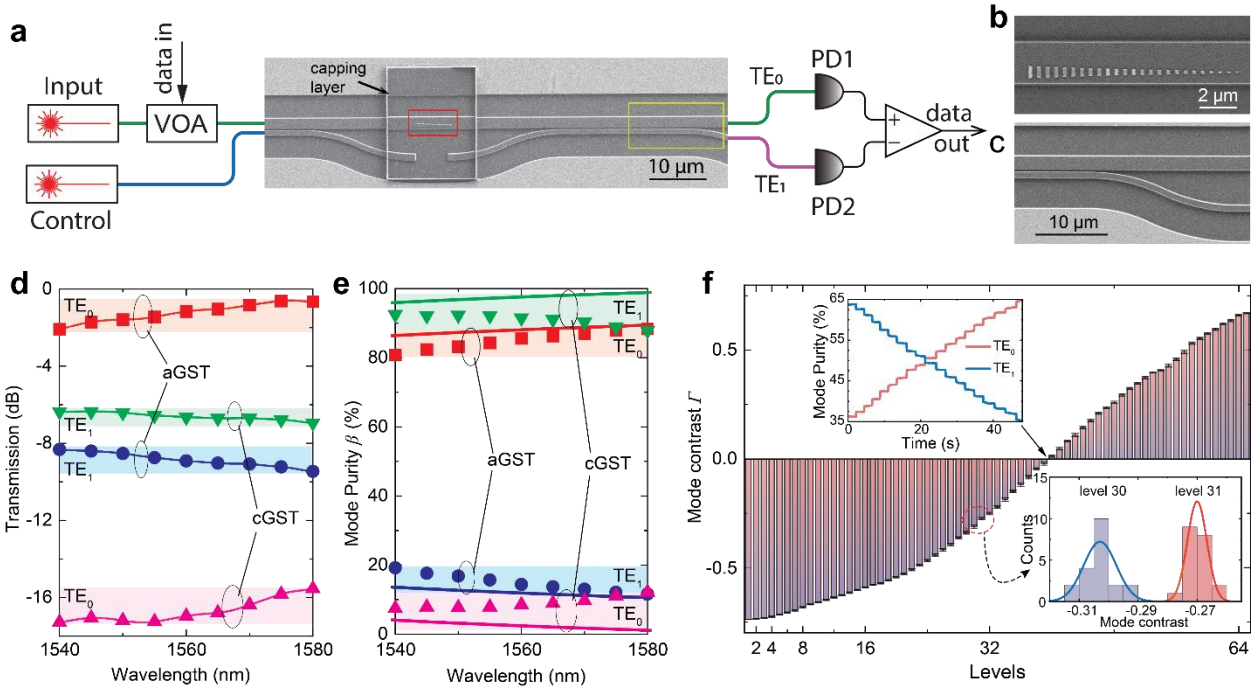

Figure 2.6 Operation of the PMMC. **a.** SEM image of the complete device and the measurement and control schematics. The complete PMMC device consists of an encapsulated GST phase gradient metasurface (red box) and mode selectors (yellow box). The white box appears from the edge of the 218 nm thick Al_2O_3 encapsulating layer. **b.** Zoomed-in SEM image of the phase-gradient metasurface on the waveguide before depositing the Al_2O_3 layer encapsulation for better imaging. **c.** Zoom-in SEM image of the TE_0/TE_1 mode selector. **d.** The transmission coefficient (insertion loss) of the devices for TE_0 and TE_1 modes and aGST and cGST phases. The transmission for the TE0 mode is switched with a high extinction ratio >16 dB or 4000%. **e.** The mode purity is controlled by the mode converter to >80% for both modes. **f.** The programable mode converter controls the mode contrast Γ at 64 distinct levels, corresponding to 6-bit programming resolution. Upper inset: zoomed-in view of the contrast levels. Lower inset: histograms of 20 programming operations to set the contrast of two adjacent levels (30 and 31). The well-separated histograms demonstrate the programming repeatability and accuracy.

The phase composition of the GST in the metasurface can be continuously tuned by partial phase transition so that the PMMC can be continuously programmed to multiple intermediate

levels of phase purity values. We program the PMMC with a sequence of 50 ns-long control pulses to "quench" the GST progressively from the fully cGST phase toward the fully aGST phase. As a result, the TE_1 mode purity β_{TE1} increases stepwise. Since the mode selector separates the two modes, we can measure their power and calculate contrast Γ. Figure 2.6f demonstrates the multi-level programmability of the PMMC, in which Γ is sequentially set to 64 distinguishable levels between -0.73 to +0.67 at 1555 nm.

To demonstrate the endurance of our device, we performed 1000 set/reset cycles of programming of our devices between levels 1 and 12. The measured mode contrast is plotted in Figure 2.7a, which shows no evidence of degradation. The slow drift is mainly due to the drift of our measurement instrument such as the alignment between the fiber array and the chip. The stability of the device was further confirmed by comparing the 64 levels of the mode contrast before and after the 1000 cycles (see Figure 2.7b).

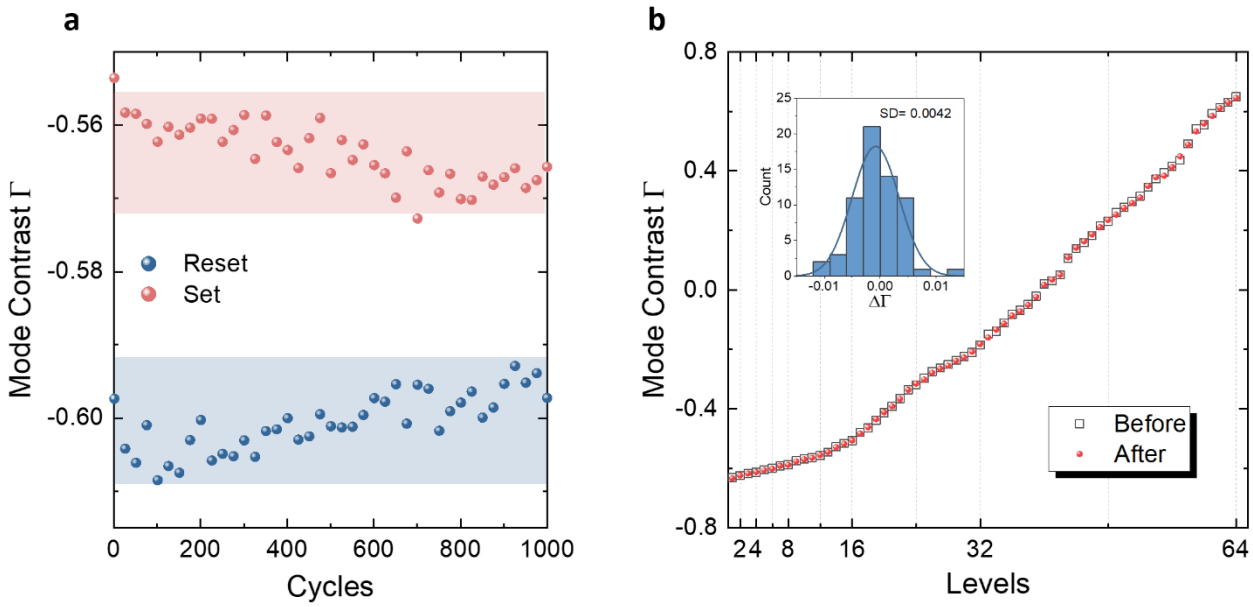

Figure 2.7 **a** The mode contrast Γ during 1000 set/reset programming cycles of our devices between levels 1 and 12. **b**. Comparison between all 64 levels of the mode contrast before and after the 1000 cycles. The inset shows the histogram. The standard deviation is only 0.0042.

2.5 Optical convolutional neural network using PMMCs

Since the theoretical range of Γ is $(-1,1)$, it is an ideal parameter to represent the elements in the matrix w, with both positive and negative values, in multiply-accumulate (MAC) operation: $x \rightarrow x \cdot w + b$, where b is the bias parameter. MAC is the constitutional step of matrix-vector multiplication (MVM) in all neural network algorithms. The PMMC allows storing w by programming Γ in the GST metasurface as a nonvolatile memory. In-memory MAC computing can be performed with the PMMC by a measurement of the transmitted power when the input data x is encoded in the power of the input optical signal. The lower inset of Figure 2.7f shows the histograms of 20 repeated programming operations to set the PMMC mode contrast at two adjacent levels (levels 30 and 31), respectively. The well-separated histograms demonstrate the device's programming resolution and accuracy. The demonstrated 64-level programmability of the PMMC corresponds to a 6-bit resolution in setting w, which is critical to the training and inference precision of the neural network [66].

We harness the PMMC's high-precision programmability and in-memory computing capability to demonstrate an optical convolutional neural network (OCNN) [67]. A typical CNN consists of an input layer and an output layer, which are connected by multiple hidden layers in between. The hidden layers usually consist of a series of convolutional layers followed by pooling layers and fully connected layers at the end. We design a prototype optical CNN using a small network of PMMCs to implement patch-kernel matrix multiplication to compute convolution. Figure 2.8a illustrates the operation principle of the OCNN for image processing, where an input grayscale image of dimensions $n \times n$ is convolved with a kernel of dimensions $k \times k$ to compute an activation map of dimension $(n–k+1) \times (n–k+1)$, assuming the convolution stride is 1. When operating the OCNN, we group the input image into $(n–k+1)^2$ patches (the shaded area in the upper panel of Figure 2.8a) with the same dimensions as the convolution kernel, k^2. Each patch corresponds to the receptive field of an element in the activation map accordingly. Thus, a convolution operation requires $(n–k+1)^2 \times k^2$ MAC operations in total, which is a high load of computation and can most benefit from optical computing's speed and energy advantages.

To compute the convolution, $(n–k+1)^2$ patch matrices of the input image are optically fed into the photonic kernel sequentially while the kernel elements, that is, the PMMCs, are programmed to fixed values. At each timeframe of the computation, the corresponding patch matrix is reshaped

into a single column of data with the length k^2. The data is input into the optical system in k^2 channels as sequences of incoherent optical pulses, whose power amplitude is controlled by a variable optical attenuator (VOA) to encode the value of each pixel value X_{ij} in the greyscale image. The corresponding element W_{ij} of the kernel matrix is programmed as the mode contrast Γ of each PMMC. The resulting transmitted power of TE_0 and TE_1 modes are then summed incoherently using two photodetectors. Their difference is calculated electronically and used in post-processing steps. As a result, the output will correspond to a time series of patch-kernel MVM with the amplitude encoding the values of the computation results, which is the activation map of convolution. Since the modal contrast Γ of our PMMCs can assume both positive and negative values, it can represent the kernel matrix elements without the need for an additional offset, which otherwise would take additional steps to set in each computation cycle.

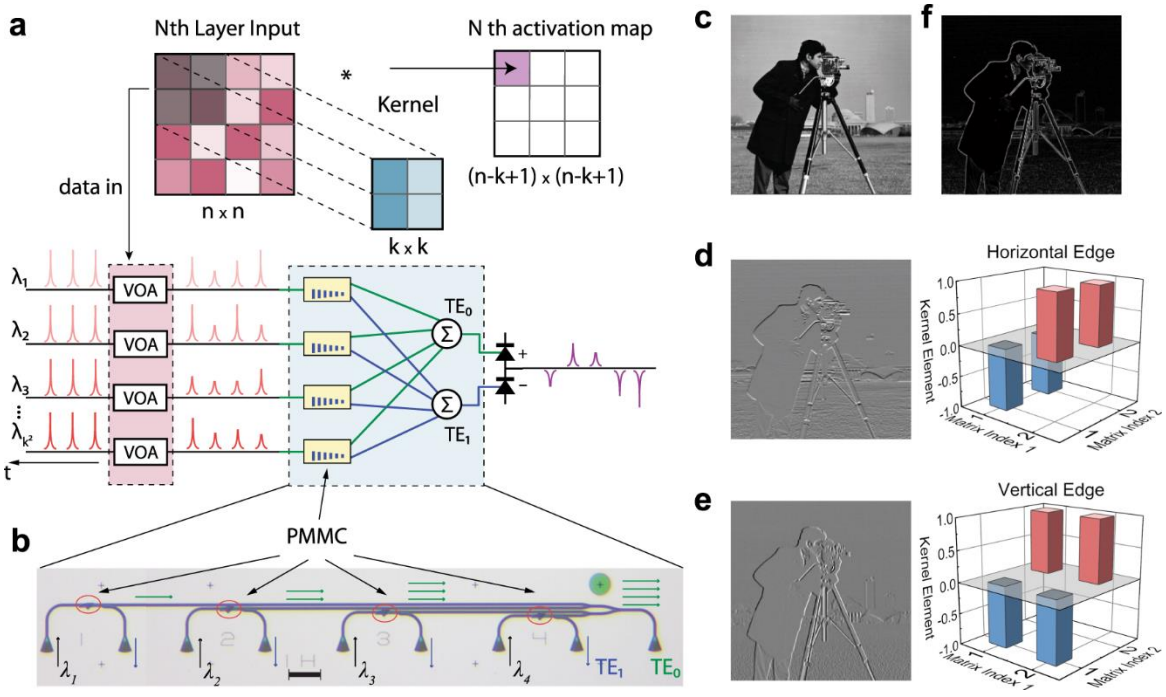

Figure 2.8 Using a PMMC array as a photonic computing core for convolutional image processing. **a.** Schematic of optical convolution for image processing. An array of k^2 PMMCs is programmed to store the kernel matrix. A patch of pixels of the image is encoded as optical pulses and input into k^2 optical channels to perform MAC operation with the kernel. The output in TE_0 and TE_1 are summed incoherently and measured with photodetectors. The activation map is represented by the mode contrast and could be both positive and negative. **b.** Optical microscope image of the

photonic core consisting of four PMMCs with four input channels. The TE_0 mode outputs are summed on-chip with Y-junctions whereas TE_1 mode outputs are summed off-chip. Optical control pulses are input using the same set of grating couplers used for the TE_1 mode detection. **c.** The greyscale image of "cameraman" (with permission from its copyright owner Massachusetts Institute of Technology) is used as the input image. **d** and **e.** Left: the raw image generated by convolution with the kernel matrix for detection of horizontal (**d**) and vertical (**e**) edges. Right: the corresponding kernel matrix for edge detection. **f.** Combined image of horizontal and vertical edge detection, highlighting all the sharp edges in the original image.

2.6 OCNN Operation procedure in steps

2.6.1 *Measurement setup for photonic convolutional tensor core*

The experimental setup used to perform the convolution operations with a 2×2 matrix is shown in Figure 2.9. The input vector is encoded as the modulated optical signal in four different wavelength channels. Another laser connected to a 1×4 optical switch is used to set the mode contrast of individual PMMC to store the value of each kernel weight element. The MVM operation is realized in the measurement of the mode contrast. The output power of the TE_0 and TE_1 modes are summed incoherently using separate photodetectors. Their difference is calculated electronically and used in post-processing steps. The TE_0 mode output from all the PMMCs is combined using on-chip Y-junctions, while the TE_1 mode output power is combined off-chip because the same ports are used to program the PMMCs optically. Because combining four incoherent sources using Y-junctions will inherently reduce the power by a factor of 4, we rescaled the measured TE_0 mode power by this factor when calculating the power differences between the two modes.

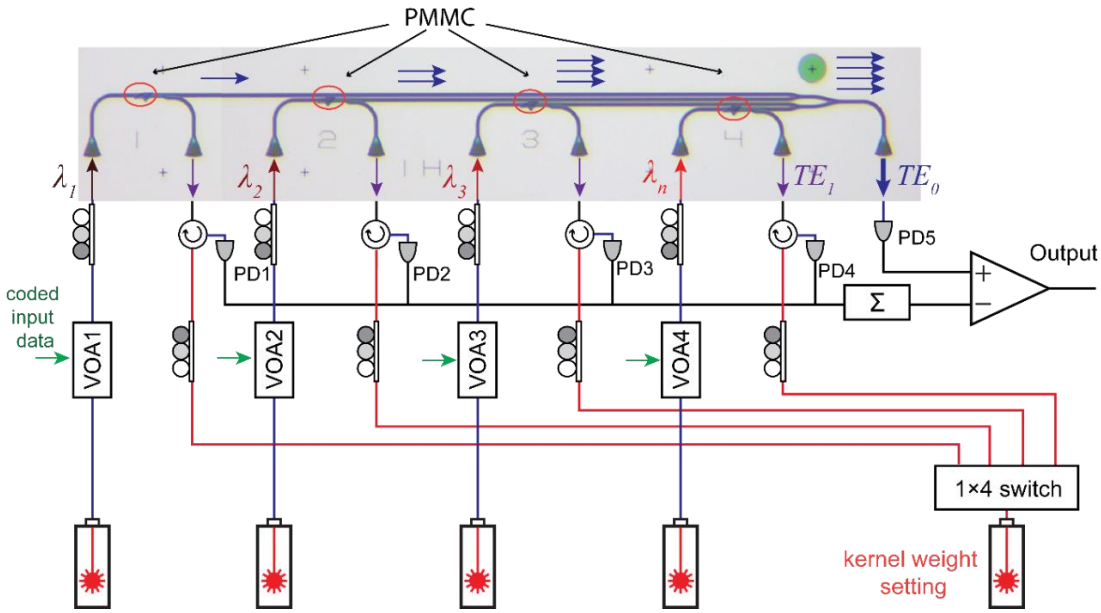

Figure 2.9 Experimental setup for convolutional MVM operation.

2.6.2 *Encoding 8-bit grayscale image in optical signal*

The first step to perform imaging processing tasks such as edge detection and pattern recognition is to encode the image from an 8-bit grayscale into the input optical signal. The 8-bit grayscale data for each pixel, represented by a decimal number between 0 and 255, is first normalized to a value in the range of [0,1]. Experimentally, this value is represented by the transmission coefficient of an optical pulse controlled by an electrical variable optical attenuator (EVOA), with no transmitted power denoting "0" (black) and maximum transmitted power denoting "1" (white). The pixel data of the image encoded in such a way is sent into the PMMCs network in a time sequence of optical pulses. Figure 2.10a shows the calibration result of the EVOA, which controls the attenuation (or transmission) of the laser pulses with an analog voltage between 0 and 5V. To test the stability and accuracy of this process, we generated 10000 random greyscale numbers with a computer and encoded them in the above process. We then measured the encoded optical pulses and compared them with the expected result, as shown in Figure 2.10b The EVOA is operated at 1 kHz. The measured result is accurate and stable compared to the expected value, and the standard deviation we calculated from the histogram is only 6×10^{-4}. Figure 2.10c and Figure 2.10d show the original image we chose to process and the image we recovered from the encoded optical signal we sent into our network. One can hardly see any difference

between these two images, demonstrating that the optical encoding process of a greyscale image is of high fidelity.

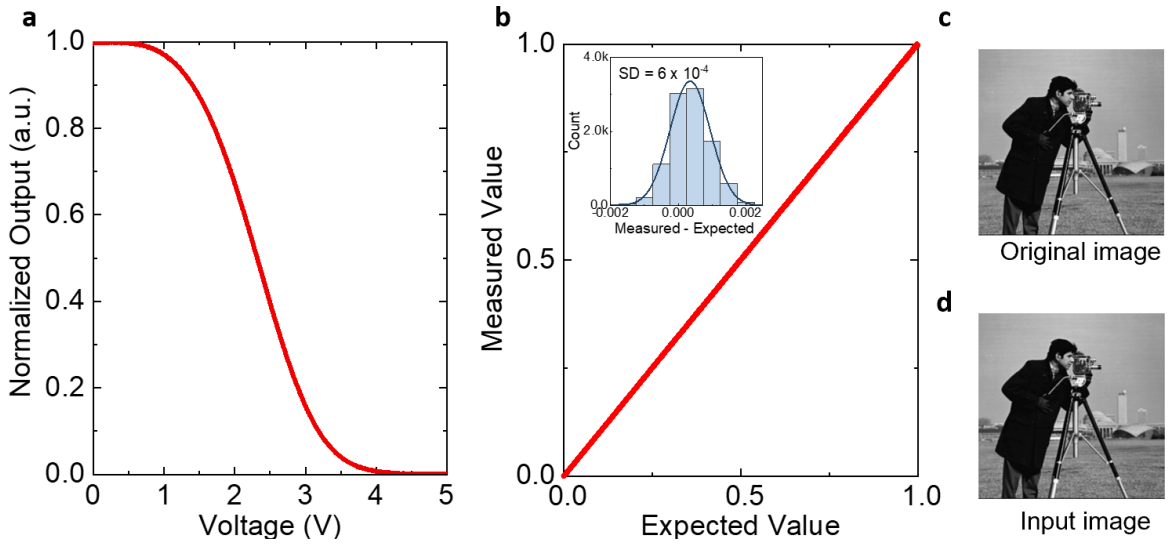

Figure 2.10 **a**. Calibration of the EVOA normalized transmission as a function of the input voltage. **b.** 10000 randomly generate grayscale numbers and corresponding measured values. The inset shows the histogram. The standard deviation is only 6×10^{-4}. **c** and **d.** The original image (**c**) as well as the input image (**d**) sent into the network recovered from the network. The original image is almost identical to the input image.

2.6.3 *Programming the PMMC matrix elements*

The next step to perform optical computing with the PMMC array is to store kernel matrices in it. A schematic of our setup for programming the PMMCs is shown in Figure 2.11. By choosing the corresponding control and probe ports, we selectively program each PMMC individually with the corresponding matrix element. The transmitted TE_1 and TE_0 mode power is measured using two photodetectors to confirm the programmed value.

The mode contrast Γ is used as the programming parameter. Without loss of generality, we describe the kernel setting procedure as follows: first, we determined the input optical power level that represents "white" pixels. Second, we calibrate all PMMCs and determine a scaling factor for each PMMC that compensates for the fabrication fluctuations and fiber alignment variations to equalize their output. This scaling factor is characteristic of each PMMC and, once calibrated, is never changed and is used in all the following measurements and operations. Third, we set the

kernel matrices to their ideal value of Γ after equalization as the programming parameters. The third step is repeated if the kernel matrices need to be reprogrammed. Figure 2.11b demonstrates how we set the kernel element in detail. To set Γ to the desired level (-0.6 in this demonstration), instead of programming with 6-bit precision all the way, we first program "coarsely" using optical pulses with a high energy to quickly approach the set level. We then program "finely" using optical pulses with a smaller energy. Benefiting from the stability and the high precision of the PMMC, the contrast will be set to the desired value precisely, as shown in Figure 2.11b. Besides, the PMMC is re-programmable by resetting it using a 500 ns long reset pulse. We here set Γ to the desired value (-0.6) 22 times continuously, the programming errors are plotted as a histogram shown in the inset. The standard deviation for our setting is only 1.20% (for 4 weights), indicating our setting process is accurate and repeatable.

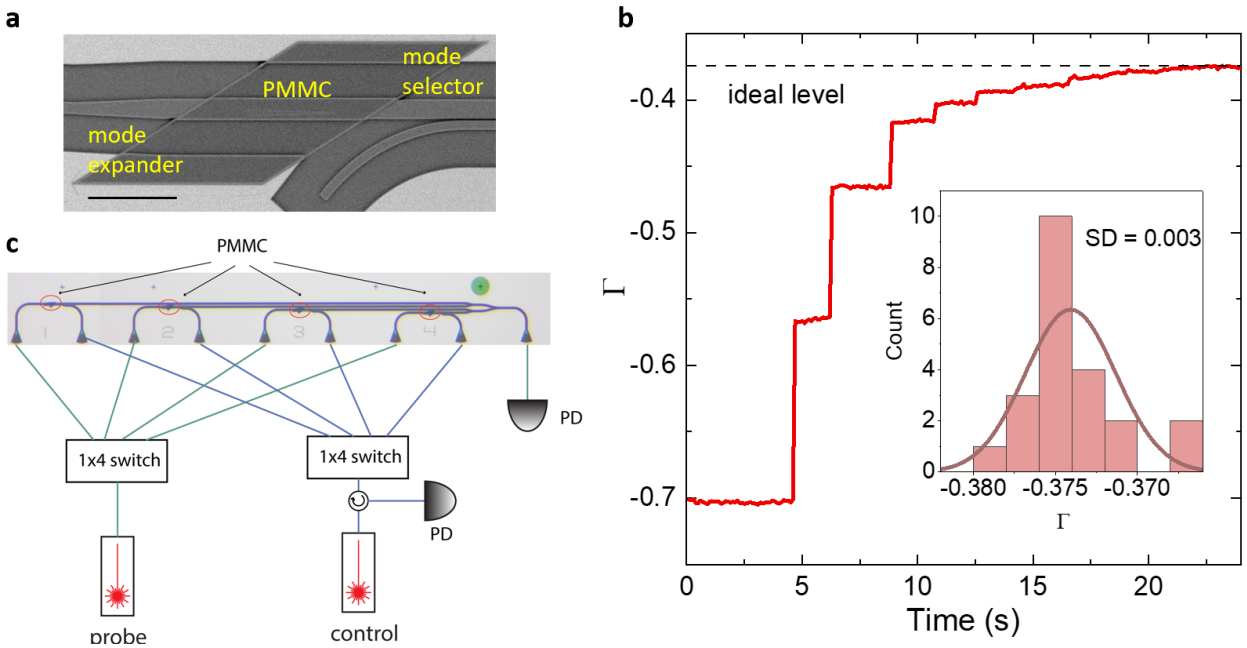

Figure 2.11 **a**. An SEM image of the PMMC. The PMMC plays the role of a convolutional kernel element in OCNN. **b.** One typical operation demonstrates how to set the mode contrast Γ to its ideal value of -0.6. Inset: the histogram calculated from repeatedly setting the kernel element to its ideal value 22 times. The standard deviation is 1.20% for four weights. **c.** The schematics for the setup used for programming the kernel elements.

2.7 Optical image processing using phase-change OCNN

2.7.1 *Convolutional Edge Detection with PMMC core*

Experimentally, we build a small-scale, four-channel system with four PMMCs to represent a 2×2 kernel matrix, as shown in the optical images in Figure 2.8b. As a demonstration, we perform the convolution of a 256×256 8-bit grayscale image of a cameraman (Figure 2.8c) to detect its edge features. As shown in Figure 2.8b, the TE_0 mode output coming from all the PMMCs is combined using on-chip Y-junctions, while the TE_1 mode output power is combined off-chip because the same ports are used to program the PMMCs optically. Because combining four incoherent sources using Y-junctions will inherently reduce the power by a factor of 4, we rescale the measured TE_0 mode power by this factor when calculating the power differences between the two modes. To detect vertical and horizontal edges, kernel matrices as in the right column of Fig 3d and e are used, and so are the PMMCs programmed. Take vertical edge detection, for example, the kernel is set to be $\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$ so to compute the discrete first-order derivative, $X_{i+1,j} + X_{i+1,j+1} - X_{i,j} - X_{i,j+1}$, where *i,* and *j* are the indices of the input image matrix. Each kernel element W_{ij} is stored as the mode contrast value Γ in the corresponding PMMC, with W_{ij} =1(-1) corresponding to the fully aGST (cGST) phase. Table 2.3 and Table 2.4 list the ideal kernel and measured kernel we used for horizontal and verticle edge detection of the image respectively.

The computed images after convolution without any post-processing are shown in the left column of Figure 2.8d, e, for horizontal and vertical edge detection, respectively. The two images are then added to produce the right image in Figure 2.8b, which highlights silhouettes of the objects with sharp edges such as the cameraman and the buildings in the original image, while suppressing smooth features such as the sky and the water. The optically computed edge detection image also agrees very well with the calculated result using conventional image processing algorithms. This result verifies the capacity and fidelity of optical convolution performed with the PMMC-based photonic kernel, which is a prerequisite for an OCNN.

Table 2.3. The ideal and experimental (rescaled) value of each kernel element for horizontal edge detection

$$K_x\,(\text{Ideal}) = \begin{bmatrix} -1 & -1 \\ +1 & +1 \end{bmatrix} = \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix}$$

	β_{TE0} (%)	β_{TE1} (%)	Γ	rescaled Γ	ideal Γ
P_1	5.21	57.25	-0.520	-0.984	-1
P_2	9.15	61.72	-0.526	-1	-1
P_3	72.43	19.15	0.533	1	1
P_4	75.94	24.06	0.519	0.963	1

Table 2.4. The ideal and experimental (rescaled) value of each kernel element for verticle edge detection

$$K_y\,(\text{Ideal}) = \begin{bmatrix} -1 & +1 \\ -1 & +1 \end{bmatrix} = \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix}$$

	β_{TE0} (%)	β_{TE1} (%)	Γ	rescaled Γ	Ideal Γ
P_1	5.89	58.18	-0.520	-0.989	-1
P_2	76.17	23.75	-0.526	0.993	1
P_3	10.57	63.92	0.533	-1	-1
P_4	76.34	23.66	0.519	1	1

2.7.2 *OCNN for Image Recognition.*

Beyond the convolution layer, the MAC computation performed with optical signals and the PMMC network can also be applied to the pooling (average pooling) and fully connected layers, where the PMMCs are used as weight banks instead, to realize a complete OCNN. In our experiment, we sequentially reuse the PMMC array in both convolution and fully connected layers to demonstrate an OCNN and perform proof-of-concept imaging recognition tasks of distinguishing handwritten numbers "1" and "2" from the MNIST database. Figure 2.12a illustrates the architecture and processes of the OCNN. The 28×28 pixels, 8-bit grayscale images of the number "1" or "2" are fed into the input layer as optical signals. The data is then convolved with two 2×2 photonic kernels $\mathbf{K_1}$ and $\mathbf{K_2}$ to generate two 27 × 27 images of activation maps. After adding a bias b_1 and applying the nonlinear ReLu function, the output images are sent to an average pooling layer with a subsampling factor of 27, which reduces the images to a 2×1 vector. This vector is then fed into the fully connected layer with a 2×2 photonic weight bank $\mathbf{K_3}$ programmed in the PMMC array, added with a bias b_2, and applied to the standard sigmoid function. The final output is a vector that gives the identified class of the input image, that is, $\begin{bmatrix}1 & 0\end{bmatrix}^T$ corresponds to the number "1" and $\begin{bmatrix}0 & 1\end{bmatrix}^T$ corresponds to the number "2". In this OCNN, the MVM computations such as the convolution and the fully connected layers are all performed optically with the PMMCs, whereas bias and nonlinear functions are realized electronically.

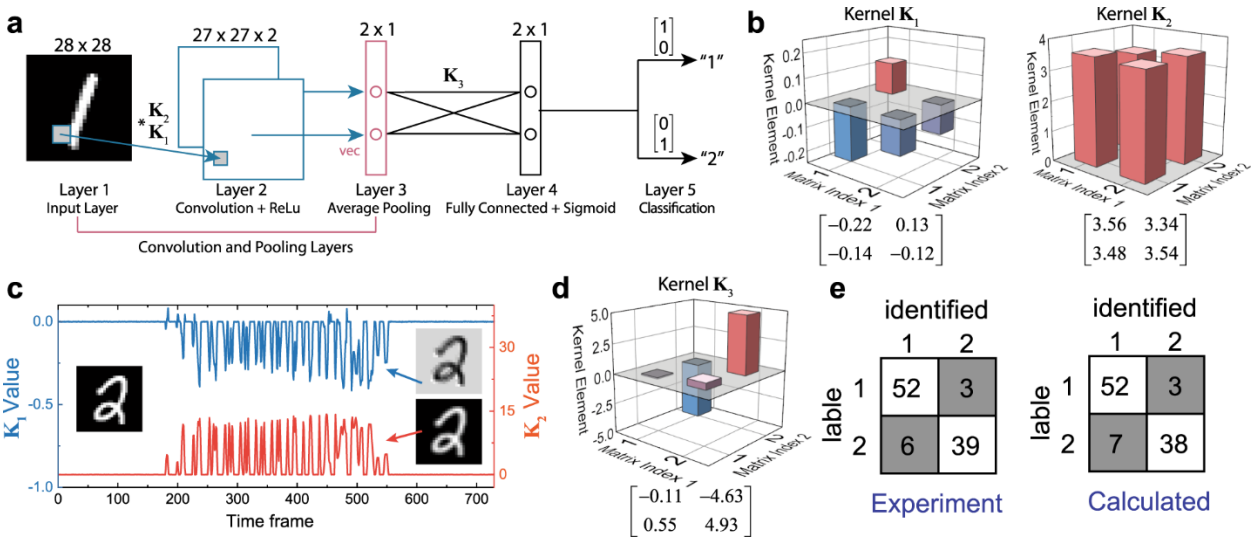

Figure 2.12 Building an optical CNN for imaging recognition. **a.** Operation procedure of using the optical CNN to recognize handwriting numbers from the MNIST database. The optical CNN consists of a convolution layer with two kernels, a pooling, and a fully connected layer. The output answers whether the input image is "1" or "2". **b.** The convolution kernel matrices $\mathbf{K_1}$ and $\mathbf{K_2}$ are generated by training the CNN. **c.** Raw output data of the convolution layer of two kernel matrices. **d.** The weight bank matrix in the fully connected layer. **e.** The recognition results from the experiment with the optical CNN (left) and calculation with a computer (right) show excellent agreement.

Before using the OCNN, we first train all the parameters in the layers with the standard back-propagation algorithm using the gradient descent method. The training set consists of 11000 images of the handwritten number "1" or "2" from MNIST training images. The training yields values for each element in the convolutional kernels and the weight bank, as shown in Figure 2.12b and Figure 2.12d. We then program the PMMC array to represent these elements. The experimentally achieved kernels $\mathbf{K_1}$, $\mathbf{K_2}$, and $\mathbf{K_3}$ are given in Table 2.5. In Figure 2.12c, we show the raw data of the convolutional activation maps encoded in a time series of optical signals, which is the output from the PMMC array after the input image convolves with the photonic kernels $\mathbf{K_1}$ and $\mathbf{K_2}$. Since each photonic processing layer results in electrical signals output from the photodetectors, electronic post-processing is performed to add bias and apply nonlinear function and pooling. The resultant data is re-coded into optical signals and fed to the next photonic layer. We evaluate the system's performance after training on a recognition test set, which consists of

100 randomly chosen "1" or "2" images (55 number “1” and 45 number “2”) from the MNIST testing image database. Figure 2.12e shows the result that our OCNN correctly identified 91 out of 100 cases (9% error rate), which compares squarely with the result of a computer (10% error rate). The slight difference is mainly caused by the small deviation of the experimentally programmed values in the matrices (**K**$_1$, **K**$_2$, and **K**$_3$) from the trained values, which occurs when the system’s conditions drift during operation. This result successfully demonstrates the OCNN’s viability and accuracy in performing standard neural network algorithms.

Table 2.5. The ideal and experimentally (rescaled) achieved kernels **K**$_1$, **K**$_2$, and **K**$_3$

$$K_1\,(\text{Ideal}) = \begin{bmatrix} P_{11} & P_{12} \\ P1_3 & P_{14} \end{bmatrix} = \begin{bmatrix} -0.2192 & 0.1369 \\ -0.1485 & -0.1272 \end{bmatrix}$$

	β_{TE0} (%)	β_{TE1} (%)	Γ	rescaled Γ	ideal Γ
P_{11}	27.55	53.36	-0.258	-0.231	-0.2192
P_{12}	52.13	37.05	0.151	0.135	0.1369
P_{13}	41.90	58.08	-0.162	-0.145	-0.1485
P_{14}	43.02	56.98	-0.140	-0.125	-0.1272

$$K_2\,(\text{Ideal}) = \begin{bmatrix} P_{21} & P_{22} \\ P_{23} & P_{24} \end{bmatrix} = \begin{bmatrix} 3.5320 & 3.5640 \\ 3.3379 & 3.5785 \end{bmatrix}$$

	β_{TE0} (%)	β_{TE1} (%)	Γ	rescaled Γ	ideal Γ
P_{21}	68.98	28.98	0.400	3.60	3.5320
P_{22}	69.13	30.56	0.389	3.50	3.5640
P_{23}	69.13	31.50	0.376	3.38	3.3379
P_{24}	69.83	30.09	0.397	3.57	3.5785

$$K_3\,(\text{Ideal}) = \begin{bmatrix} P_{31} & P_{32} \\ P_{33} & P_{34} \end{bmatrix} = \begin{bmatrix} -0.1415 & -4.8810 \\ 0.1620 & 4.8829 \end{bmatrix}$$

	β_{TE0} (%)	β_{TE1} (%)	Γ	rescaled Γ	ideal Γ
P_{31}	32.88	33.46	-0.006	-0.122	-0.1415
P_{32}	24.53	47.90	-0.234	-4.739	-4.8810
P_{33}	46.05	43.45	0.026	0.526	0.1620
P_{34}	62.47	37.53	0.249	5.04	4.8829

2.8 The perspective of scalability— crossbar array architecture

2.8.1 *Work principle of the crossbar array*

The network architecture can be scaled up using a crossbar array architecture as shown in Figure 2.13.

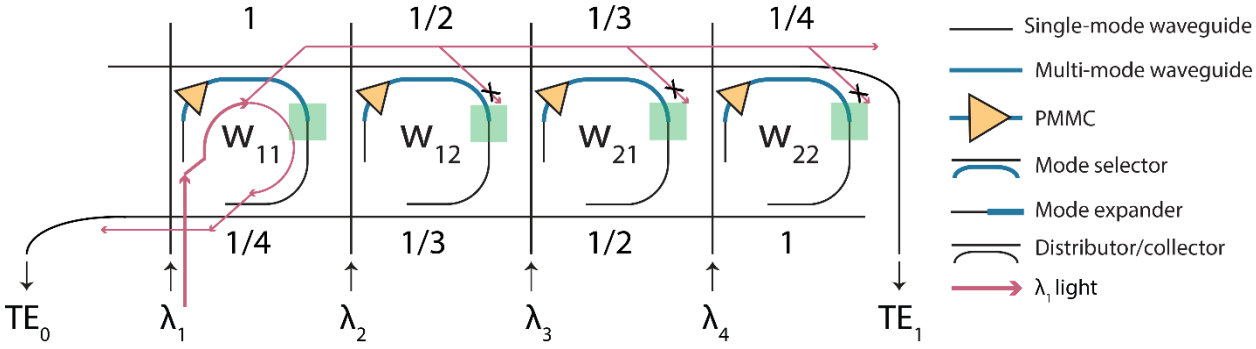

Figure 2.13 Schematic of a photonic crossbar array architecture used to perform optical convolution operation with a 2×2 kernel matrix. The arrows denote the detailed optical path for the first input channel (λ_1). The cross-coupling ratios for each mode selector and mode expander are labeled, respectively.

All the TE_1 output from each of the 4 units is combined incoherently in the top horizontal waveguide and summed at the output. Likewise, all the TE_0 outputs are combined in the lower horizontal waveguide and summed. It is thus important that the output from each unit is combined

with the same weight. This is achieved by the arrangement of the directional couplers between each unit and the bus waveguide. We elaborate on the steps below:

The input TE_0 mode in the first channel with wavelength λ_1 passes through the first unit of PMMC, which corresponds to the kernel element w_{11} and is partially converted to TE_1 mode based on the value of w_{11}. So this is doing the part of the multiplication of $x_1 \cdot w_{11}$ (after taking a difference from the TE_0 mode). The TE_1 mode component couples to the top bus waveguide with a 100% coupling ratio and then passes by the next three units before reaching the output. In the second unit with kernel element w_{12}, the coupling ratio between the PMMC and the bus waveguide needs to be designed to be 1/2, so the through port efficiency of the bus waveguide is also 1/2. For the third unit with kernel element w_{13}, the PMMC to bus coupling efficiency should be 1/3 and the through port efficiency of the bus is 2/3. Finally, in the fourth unit, the two coupling ratios should be 1/4 and 3/4, respectively. Therefore, the overall collective efficiency of the TE_1 mode output from the first unit will be $1 \times 1/2 \times 2/3 \times 3/4 = 1/4$. Similarly, the overall collective efficiency of the TE_1 mode output from the second unit will be $1/2 \times 2/3 \times 3/4 = 1/4$. For the third and fourth units, the overall efficiency will be the same. The collective efficiencies for the TE_0 mode power from each unit to the bottom bus waveguide are designed in the same way to be 1/4. These coupling efficiencies are denoted in the figure above. Also, note that the TE_1 light left in the multimode waveguide will be filtered out by the mode expander (the green box) so will not be collected by the lower bus waveguide.

2.8.2 *Mapping convolutions to photonic MAC operations*

To further scale up our system to realize large convolution kernels in parallel, a photonic crossbar array architecture as sketched in Figure 2.14 can be used. The photonic crossbar array can perform a general MVM operation: $\boldsymbol{y}=\mathbf{W}\cdot\boldsymbol{x}$, where $\boldsymbol{x}=[x_1, x_2, \ldots, x_n]^T$ is the input vector, the $\mathbf{W}$ is the $m \times n$ matrix represents m convolutional kernel matrices in parallel $\boldsymbol{y}=[y_1, y_2, \ldots, y_m]^T$ is the output vector. The horizontal and vertical single-mode waveguide separates the whole structure into $m \times n$ subunits. The input vector x is encoded through a group of wavelength $[\lambda_1, \lambda_2, \ldots, \lambda_n]$, with the value of each element x_i represented by the input power, and sent into n horizontal bus waveguides. A directional coupler-based power distributor is used to distribute the power carried by the bus waveguide evenly into m subunits. These m subunits in a row of the crossbar array

correspond to a column in the matrix **W**. The 2*m* vertical single-mode waveguides are used as power collectors to collect the light after leaving each PMMC. The TE_0 and TE_1 lights collected and summed from the same row in the matrix **W** (the same column in the crossbar array) will be grouped in pairs and measured with balanced photodetectors to determine the contrast. Both the photonic convolution and the MVM operations are based on the photonic MAC operations, the *m* convolution kernels with the dimensions of $\sqrt{n}\times\sqrt{n}$ thus can be mapped to the **W** matrix. The operation volume can be further extended through the wavelength-division multiplexing (WDM) method. For example, if *g* groups of the wavelength are implemented parallelly into the crossbar network with *n* wavelength per group, the numbers of the MAC operation performed can increase by *g* times, *g* × *m* × n MAC operations or *g* × *m* convolutions can be applied in parallel.

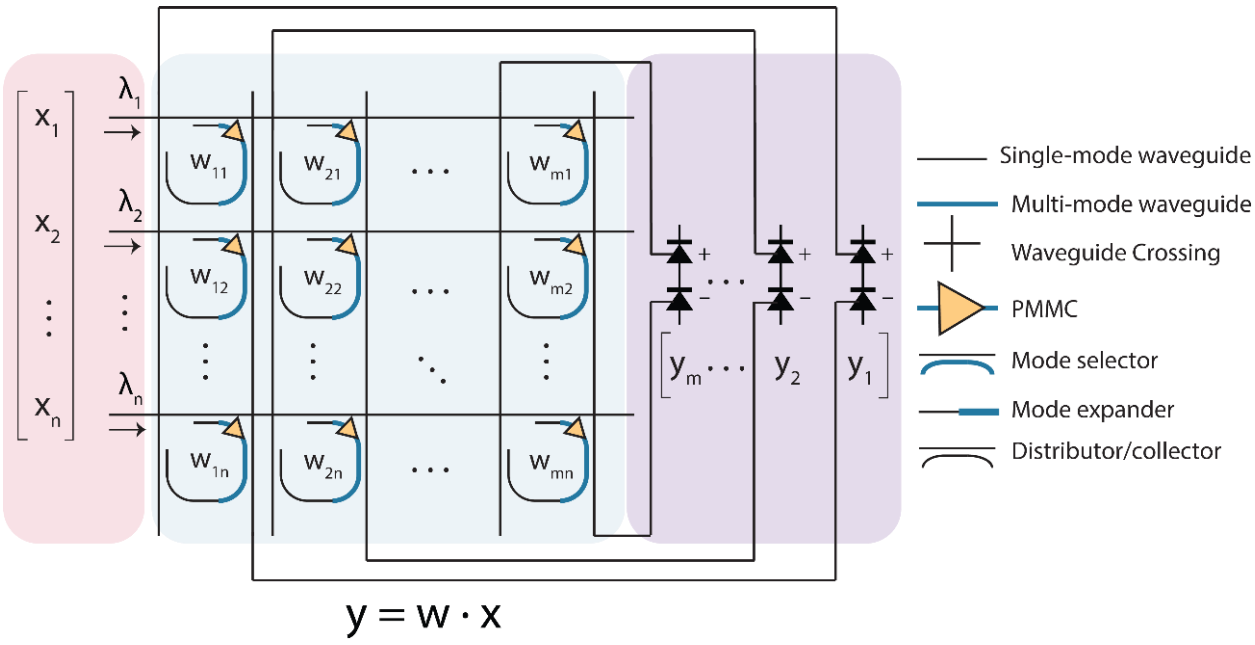

Figure 2.14 Building The schematic of the photonic crossbar architecture used to perform MVM operation on a larger scale. Horizontal and vertical waveguides separate the network into *m* × *n* subunits

2.8.3 *The limiting factors of scalability and comparison with the state-of-the-art commercial microprocessors*

The insertion loss is for the TE_1 mode when the GST is in the cGST phase (TE_1 mode at aGST phase and TE_0 mode at cGST mode is supposed to be suppressed). As shown in, this insertion loss Figure 2.6d (green triangle symbols) of our PMMC device is ~7 dB in the measured wavelength range. Although lowering this insertion loss is beneficial to scaling, it is not the bottleneck for a crossbar array architecture that is commonly used. This is because the optical signal in each channel only passes one PMMC once to perform the multiplication: $w_{ij} \cdot x_i$. Rather, in a large network, the optical power loss will be dominated by the directional couplers. To equalize the output from n units, the directional couplers that are used to collect the output from each unit to the bus waveguide need to be designed to achieve a $1/n$ weight for each unit. Therefore, when $n>5$, this loss will be larger than the 7 dB insertion loss of the PMMC.

With these numbers, we notice that a crossbar array size is limited by the optical loss, assuming no optical amplifiers are used to boost up the power. If we assume the input optical power for a single wavelength channel is at a moderate level of 10 mW (10 dBm), and a 10 GHz bandwidth photoreceiver with noise-equivalent power of 10 $pW \cdot Hz^{-1/2}$ (e.g. RXM10AF from Thorlabs) is used. So 40 dB total insertion loss is allowed. This means n can be as large as 2000, considering the 7 dB loss of a PMMC (40-7 dB=33 dB, approximately 1/2000). This is already a very large network, corresponding to a kernel matrix of ~45×45, or 45 5×9 kernel matrices in parallel. To go beyond this size limitation imposed by device insertion loss, optical amplifiers will need to be inserted into the network. It is possible to integrate semiconductor optical amplifiers (SOA) in the photonic network. The amplifiers, however, will increase the noise figure and negatively impact the accuracy of the network.

The photonic crossbar array architecture with WDM can fully utilize the intrinsic parallelism of photonic systems. We can estimate and compare the expected computation performance, including speed (Tera-Operations per second (TOPS)), operation power, clocks, MAC sizes, computing density, and energy efficiency (TOPs/W), of the PNN with commercial GPUs commonly used in as accelerators for neural network-based AI applications.

We assume an array size of 32×32, an operation speed of 10 Gbits/sec in data rate, and 16 wavelengths used in WDM. All of these values are quite moderate compared with state-of-the-art optical communication technology. Since our PMMC device is very compact, its footprint

(including the mode selector) is only 80×20 μm^2, a 32×32 array will have an area of less than 2 mm^2. With 10 Gbit/sec operation frequency and 16 wavelengths, the network's computation speed will be 10G/sec×16×32×32=164 TOPS. Its areal computing density will thus be 82 TOPS/mm^2. These values compare very favorably with the current digital computing technology for neural networks, such as GPUs and TPUs. In terms of computation density (TOPS/mm^2), the photonic architecture is 800× higher than GPUs, 60× higher than TPUs, and 20× higher than the emerging memristor processors (see Table 2.6).

Table 2.6. Comparison of the projected performance parameters of a photonic CNN with the commercial electronic processors.

Processor	**Format**	**TOPS**	**Clock [GHz]**	**MAC Size**	**MAC Area (estimated) [mm^2]**	**TOPS/mm^2**
Nvidia GPU P40	INT 8	48	1.3GHz	-	471	0.1
Nvidia GPU V100	INT 8	62.8	1.3GHz	-	815	0.08
Google TPU	INT 8	90	700MHz	256×256	72	1.25
IBM (Memristor)	INT 8	1.4	4.2MHz	512×512	0.4	3.6
Photonic CNN (16-λ)	**INT 6**	**164**	**10GHz**	**32×32**	**2**	**82**

2.9 Conclusion and outlook

In summary, we have demonstrated a compact programmable waveguide mode converter using a GST-based phase-gradient metasurface with high programming resolution, efficiency, and broadband operation. We have built a photonic kernel based on an array of such PMMC devices and implemented an optical convolutional neural network to perform image processing and recognition tasks. Our results show that phase-change photonic devices, such as the PMMC demonstrated here, can enable robust and flexible programmability and realize a plethora of unique optical functionalities that are scalable for large-scale optical computing and neuromorphic photonics. Although optical computation in this work is performed at a low speed of ~1 kHz by using low-speed VOAs to encode data into optical signals, state-of-the-art integrated photonic transmitters and photodetectors can drive the system at a speed of many 10s of Gbits/sec [27,29]. Using wavelength division multiplexing (WDM) can further increase the number of parallel computations. The 2×2 array prototype system demonstrated in this work performs optical computation incoherently in broadband. It can be scaled up toward a large network using a photonic crossbar array architecture and compares favorably with other photonic computing schemes using coherent methods[30] or optical resonators. The feasible size ($n \times m$) of such crossbar arrays will not be limited by the insertion loss of the PMMC (~7 dB for TE_1 mode); rather it will be limited by the directional couplers with coupling efficiency of $1/n$, as is needed to equally combine signals from n units. Scaling up to a large network thus faces the challenge of diminishing optical power unless with on-chip optical amplification, which is not yet available. Still, an OCNN system using the PMMC device can afford an extremely high areal computing density (defined as MAC operations per time per unit area) because of its compact footprint of ~80×20 μm. For example, assuming a moderate data rate of 10 Gbits/sec and 4 WDM wavelengths in parallel per channel, the computing density will reach an upper bound value of 25 TOPS/mm^2 (Tera-Operations per second per mm^2), which is significantly higher than that of digital electronic accelerators such as GPUs and tensor processing units (TPUs). Using silicon instead of silicon nitride can further reduce the device footprint to increase the computing density. Besides MAC operation, the equally important computing processes of applying nonlinear functions and pooling can also be achieved optically by using elements such as nonlinear optical resonators, modulators, and amplifiers. Alternatively, a hybrid photonic-electronic system may optimally balance the

energy-efficiency and speed advantages of photonic systems, while realizing flexible non-linearity, connectivity, and training precision using microelectronics [28]. With these advances and after overcoming the scaling challenge, the photonic neural network accelerator will be very promising for AI in data centers where massive optical interconnects have already been deployed.

Chapter 3. Harnessing Optoelectronic Noises in a Photonic Generative Network

The current rate of improvement in digital electronics' energy efficiency is lagging behind the fast-growing computational load spurred by the widespread implementation of large-scale artificial neural networks for machine learning and artificial intelligence [68–70]. Because of its significant advantages in power efficiency, communication bandwidth, and parallelism, analog optical computing based on integrated optoelectronic processors is once again brought into focus as hardware accelerators for neural networks [71–74]. Photonic neural networks reported to date [6,7,63,74] are predominantly hybrid optoelectronic networks, in which the photonic components are used for linear multiplication and interconnect while nonlinear functions and feedback control are implemented electronically. Compared to electronic neural networks using digital processors, photonic neural networks have higher inaccuracy and error rates due to their analog nature and the abundance of optoelectronic noises in the hardware. The accumulation of computational errors in large-scale photonic neural networks could severely impair their performance, limiting the computation effectiveness and scalability [75–77]. Although several offline noise-aware training schemes, including injecting noises to layer inputs [76,78], synaptic weights [75,79], and pre-activations [80,81], have been proposed to mitigate analog hardware non-idealities, those schemes only address discriminative models [77,82–84]. In another study, a diffractive optics-based network is trained with carefully drafted parametric randomness to be robust against optical non-idealities [85,86]. Noise in the analog hardware has also been utilized to facilitate various machine-learning algorithms [87–90]. In contrast to discriminative models, generative neural network models can automatically discover and learn regularities or patterns from the training data to generate plausible new instances [91–93]. So far, a photonic generative network has not been reported, and the corresponding noise mitigation strategies have not been explored.

In this chapter, we demonstrate a generative network (GAN) based on a photonic computing core with the same design [63] as we used in Chapter 2. The photonic generative network is combined with a discriminator to realize a GAN that is trained to generate handwritten numbers. We show that the photonic GAN can harness and mitigate optoelectronic noises and errors in three ways. First, we utilize the amplified spontaneous emission (ASE) noise to realize an optical true

random number generator (RNG), which is used as the input to the GAN. This optical RNG efficiently generates random numbers at high speed in multiple wavelength channels by slicing the ASE spectrum [94,95]. Second, we analyze error sources originating from the components in the photonic GAN and propose noise-aware training approaches by augmenting noises during the training process, which improves the network's performance and robustness. Third, we validate the training approaches through experiment and simulation, and demonstrate that the photonic GAN can benefit from the inevitable random errors in practical implementation. Surprisingly, the images generated by non-ideal photonic hardware show even higher quality than those by ideal, errorless counterparts (*i.e.*, software baseline). Our results demonstrate the feasibility and resilience of more complex photonic GANs using non-ideal optoelectronic hardware. Since the proposed noise-aware training approaches are generic, they can be applied to various types of optoelectronic neuromorphic computing hardware.

3.1 Photonic generative networks in a GAN architecture

A GAN network consists of two sub-neural network models (see Figure 3.1a), a generator, and a discriminator [96–98]. These two models compete against each other in a zero-sum game: the discriminator strives to distinguish the instances produced by the generator (labeled as the "fake" instances) from the real instances in the training dataset (labeled as the "real" instances); the generator aims to fool the discriminator by producing novel instances that imitate the real instances. The competition drives both networks to improve their capabilities until an equilibrium state is reached, *i.e.*, when the "fake" instances are indistinguishable from the "real" instances by the discriminator, so the generator is deemed well-trained to generate plausible new instances. In this work, we design a prototype photonic generator to produce images of the handwritten number "7" based on a noise-aware offline training configuration: we first train the generator model on a computer and implement it on the photonic platform (Figure 3.1b). In this chapter, we only focus on realizing the photonic generator since the photonic discriminator is typically a convolutional neural network which has been demonstrated in Chapter 2. As shown in Figure 3.1c, in each layer of the generator, the input data is encoded in the power of the optical signals through multiple wavelength channels, processed by the PMMC photonic tensor core (Figure 3.1d), in which the

kernel matrices are stored. The results are detected by the photodetector arrays. Electronic post-processing is then performed to apply nonlinear functions. The results are re-encoded into the optical signals and relayed to the next photonic layer. In such an optical network, various noises, including optical and electrical noises of the optical sources, modulators, and photodetectors, accumulate through the processes of programming (*i.e.*, writing) the kernel matrices, data encoding, and data transferring (*i.e.*, reading) between the layers of the network. Because in realistic experiments, the computation errors stem from various optoelectronic noises in the system, we use the terms noise and error interchangeably in this chapter.

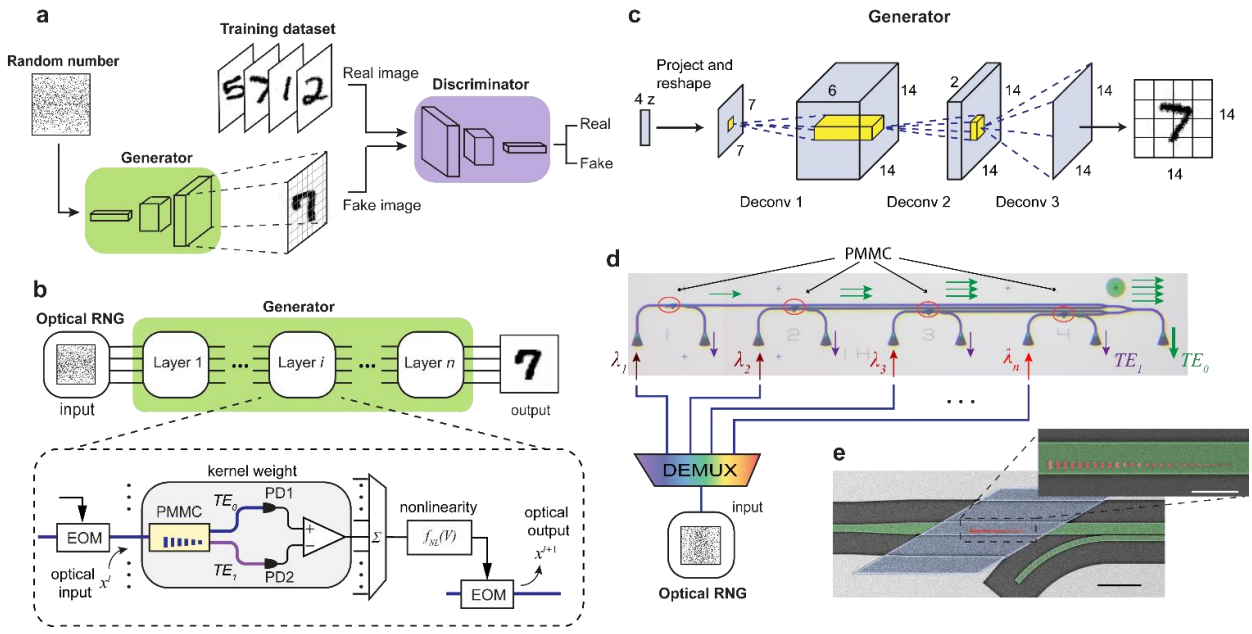

Figure 3.1 Photonic GAN network with optoelectronic noises. **a.** A GAN architecture is composed of two sub-network models, a generator and a discriminator. The generator competes with the discriminator during training and produces new instances after it is trained. **b.** The offline noise-aware training and inference processes flow of the generator. The process of mapping the trained weight to the hardware during implementation inevitably introduces optoelectronic noise. **c.** Decomposition of the generator into individual layers. In each layer, the input signals pass through the photonic tensor core and are converted to the electrical domain by photodetectors (PDs). After post-processing, the data is converted back into the optical domain and transferred to the next layer. **d.** Optical microscopic image of the photonic tensor core consisting of four input channels. The random noise is fed into the photonic tensor core through O/E and E/O conversion in our

experiment. Potentially, the optical noise can be directly sent into the tensor core using WDM schemes. **e.** The detailed false-colored SEM image of the photonic tensor core. This device is in the same design as shown in Chapter 2. The Si_3N_4 waveguide, the GST metasurface, and the Al_2O_3 protection layer are colored green, red, and blue, respectively. Scale bar: 10 μm. Inset: the zoomed-in SEM image of the phase-gradient metasurface on the waveguide. Scale bar: 2 μm.

There are two key components of the photonics generator: the optical random number generator and the photonic tensor core. We note that the photonic tensor core has the same design as we demonstrated in Chapter 2. In Section 3.2 and Section 3.3, we show the characterization of the ORNG part and tensor core part (through MVM error analysis) respectively.

3.2 Optical random number generator

3.2.1 *Theory and process flow of the optical random number generator*

One key component of the photonic generator is the optical RNG that produces the random input. Here, we followed the method developed by Williams *et al.* [94] to generate random signals from the ASE noise, as shown in Figure 3.2. The input signal $u(t)$ is the optical random noise produced from the amplified spontaneous emission (ASE) noise of the EDFA, which has a broadband power spectral density (PSD) $S_{in}(f)$. $u(t)$ is first spectrally filtered after passing an optical band-pass filter with the frequency response $H_{BP}(f)$. Within the narrow passband B_{BP}, the PSD is approximated as a constant S_{in} so the PSD after the filter is $S_{in}|H_{BP}(f)|^2$. The filtered optical signal is then detected by a square-law photodetector. The generated electrical signal passes an electrical low-pass filter with the frequency response $H_{LP}(f)$ and bandwidth B_{LP}, generating noisy baseband electrical voltage signals from the beating between different optical frequency components, referred to as "ASE-ASE beat noises." In practice, the photodetector plays the role of both the power meter and the low-pass filter. Here, we assume the responsivity R (in V/mW) and the gain G of the photodetector are constants. The frequency responses of the band-pass and low-pass filters are expressed as below:

$$\left|H_{BP}(f)\right|^2 = \exp\left[-(4ln2)\frac{(f-f_0)^2}{B_{BP}^2}\right] \tag{3.1}$$

$$\left|H_{LP}(f)\right|^2 = \exp\left[-(ln2)\frac{f^2}{B_{LP}^2}\right] \tag{3.2}$$

where f_0 is the center frequency of the band-pass filter. As a result, the PSD of the electrical noise S_{noise} after the DC block is given by the integration of optical noise power and the photodetector response and has a Gaussian profile:

$$\begin{aligned} S_{noise}(f) &= R^2G^2S_{in}^2\left|H_{LP}(f)\right|^2\int\left|H_{BP}(f')H_{BP}(f'+f)\right|^2 df' \\ &= R^2G^2S_{in}^2B_{BP}\sqrt{\frac{\pi}{8ln2}}\exp\left[-(ln2)\left(\frac{1}{B_{LP}^2}+\frac{2}{B_{BP}^2}\right)f^2\right] \end{aligned} \quad . \tag{3.3}$$

The corresponding voltage variance at the final output thus is given by:

$$\begin{aligned} \sigma_{noise}^2 &= \int S_{noise}(f)\,df = R^2G^2S_{in}^2\int\int\left|H_{LP}(f)H_{BP}(f')H_{BP}(f'+f)\right|^2 df'df \\ &= R^2G^2S_{in}^2B_{BP}\sqrt{\frac{\pi}{4ln2}}\left(1+\frac{B_{BP}^2}{2B_{LP}^2}\right)^{-1/2} \end{aligned} \quad . \tag{3.4}$$

In practice, we adjust the EDFA gain to control the mean power of the ASE noise (DC component), $S_{in}H_{LP}(0)\int\left|H_{BP}(f)\right|^2 df = S_{in}B_{BP}\sqrt{\frac{\pi}{4ln2}}$, and make sure that it will not saturate the photodetector.

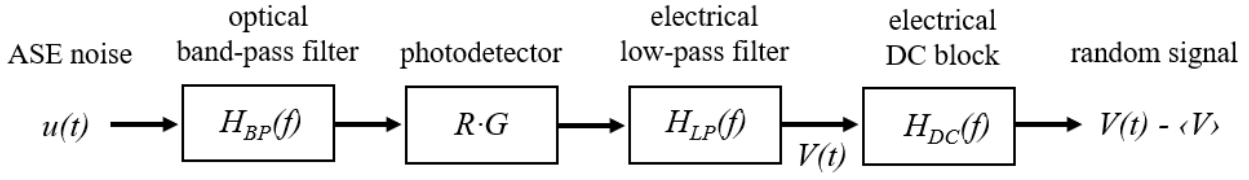

Figure 3.2 The block diagram showing the key components and the process flow used to generate random numbers from the ASE noise.

3.2.2 *Characterization of optical random number generator using ASE-ASE beat noise*

To realize an ORNG experimentally, we utilize the ASE noise from the erbium-doped fiber amplifiers (EDFA), the ubiquitous noise source in fiber-optic communication systems, to generate random optical signals at high rates in four parallel channels as shown schematically in Figure 3.3a. The EDFA generates the ASE noise signal in a very broadband optical bandwidth. Based on the theory described in Section 3.2.1, here we used wavelength division demultiplexers (DEMUX) as the optical bandpass filter which owns four independent wavelength pass channels. The ASE noise is first filtered with wavelength DEMUX and then detected with photodetectors. The generated baseband electrical currents due to beating between different frequency components are the so-called "ASE-ASE beat noise" [99,100]. The DC photocurrent is filtered by a DC block, passing only the stochastic photocurrent variances to a sampling oscilloscope to generate random numbers. Figure 3.3c and Figure 3.3d plot a representative trace of the random numbers (in voltage) generated in a single WDM channel, and the statistical histogram respectively. The probability density function is well-approximated by a zero-mean Gaussian distribution with a standard deviation (SD) of 0.2 V (*i.e.*, $N(0, 0.2)$). We further calculate the correlation coefficient of an $N=5\times10^4$-number long sequence (Figure 3.3e), which reaches the limit of $1/\sqrt{N}$ (red line in Figure 3.3e), proving the randomness of the number sequence. Because of the limited size of the photonic tensor core, we need to measure and record the random numbers from the RNG and repeatedly input them to the generator during the experiment (see Figure 3.1d). In future full-scale systems, the filtered ASE noise can be directly used as random optical inputs to the GAN without electrical sampling (the dashed box in Figure 3.1d) and detected after the first layer of the network is performed.

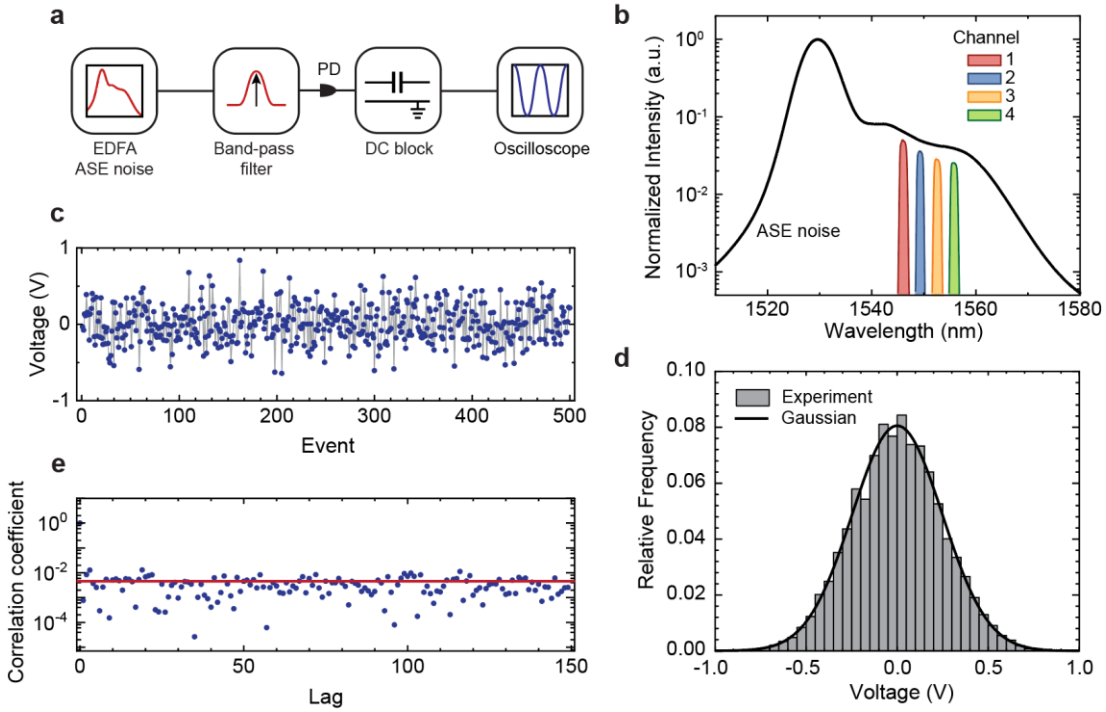

Figure 3.3 **a.** Schematic of the optical RNG. The ASE noise is spectrally sliced into four wavelength channels using a wavelength division demultiplexer and then detected with photodetectors. After a DC block, the random electrical signals are sampled by an oscilloscope. **b.** The spectra of the overall ASE noise before (black) and after were filtered, which allows the generation of random signals in four independent channels in parallel. **c** and **d.** A representative event trace (**c**) and statistical histograms (**d**) of the generated random numbers. The generated random number follows the Gaussian distribution. **e.** Correlation coefficient as a function of lag for the random number sequence. A random number sequence with length $N = 5\times10^4$ has a correlation coefficient (blue dots) around the lower limit $1/\sqrt{N}$ (red line).

3.3 MVM error analysis for PMMC-based photonic tensor core

3.3.1 *Error sources contributed to MVM errors*

The other key component of the photonic generator is the photonic tensor core, which optically performs matrix-vector multiplication. Recall the schematic of one PMMC kernel element of the core that MAC: $x \rightarrow x \cdot w + b$, the fundamental operation of MVM. The input vector element x is encoded in the power of the input optical signal. The corresponding kernel element weight w is

represented using the TE_0/TE_1 mode contrast $\Gamma = \beta_{TE0} - \beta_{TE1}$ at multiple intermediate levels between [-1..1], where $\beta_{TE0\,(TE1)} = P_{TE0\,(TE1)} \big/ \left(P_{TE0} + P_{TE1} \right)$ is the mode purity, and P_{TE0} (P_{TE1}) is the power of the TE_0 (TE_1) mode component in the waveguide. Thus, the MAC computation is simplified to an incoherent optical transmission measurement and can be performed over a broad bandwidth (Figure 3.1c).

The key to generating high-quality handwritten number images is to perform accurate matrix-vector multiplication (MVM) operations, $\boldsymbol{y}^l = \mathbf{W}^l \cdot \boldsymbol{x}^l$, where $\mathbf{y}^l$ is the output matrix, $\mathbf{W}^l$ is the kernel matrix, and $\mathbf{y}^l$ is the input vector of the layer l. In practice, the noise is inevitably introduced through non-ideal pieces of equipment, leading to errors on both input vector $\boldsymbol{x}^l$ and the kernel matrix $\mathbf{W}^l$, the realistic MVM operation will give the results:

$$\begin{aligned} \boldsymbol{y}^l{}' &= \boldsymbol{y}^l + \Delta\boldsymbol{y}^l = \left(\mathbf{W}^l + \Delta\mathbf{W}^l \right) \cdot \left(\boldsymbol{x}^l + \Delta\boldsymbol{x}^l \right) \\ &\approx \mathbf{W}^l \cdot \boldsymbol{x}^l + \Delta\mathbf{W}^l \cdot \boldsymbol{x}^l + \mathbf{W}^l \cdot \Delta\boldsymbol{x}^l \end{aligned} \tag{3.5}$$

where $\Delta\boldsymbol{y}^l$, $\Delta\mathbf{W}^l$, $\Delta\boldsymbol{x}^l$ are the errors for the corresponding elements. $\Delta\boldsymbol{x}^l$ is mainly caused by the inaccurate response of the EOM at the input (input encoding error). $\Delta\mathbf{W}^l$ is caused by mode contrast setting error, $\Delta\Gamma^l$, which includes short-term mode contrast setting inaccuracy $\delta\Gamma$ (write error) and the long-term measurement fluctuations (read error) such as the drift of the measurement setup over time.

3.3.2 *Characterize the input encoding error*

To analyze the error introduced during the encoding process of the input sequence, we commence by generating a 10000-unit-long random sequence adhering to a Gaussian distribution of $N(0, 0.2)$. This random sequence is subsequently divided into two separate subsequences: one comprising solely positive inputs and the other consisting of exclusively negative inputs.

Independently, we encode the amplitude of the input optical power in each of these two subsequences using an electro-optic variable attenuator. Following the amplitude modulation, we directly measure the power of the encoded signal. Subsequently, we compute the difference

between the two sequences to derive the desired inputs. This comparative analysis allows us to characterize the discrepancy induced during the encoding process.

Figure 3.4a plots a 100-events long measured optical signal trace chosen among the whole length of the sequences the input to the first layer obtained from measurement with its corresponding ideal value. The circles and the solid dots are the ideal value and the corresponding measured value respectively. Upon analysis, we observe a close match between the measured input values and the ideal input values that were initially set. This indicates a favorable alignment between the expected (ideal) and actual (measured) input values, signifying an accurate encoding process of the input values. We further plot the histogram of the input error in Figure 3.4b. Here, the input error is defined as $|\Delta \boldsymbol{x}| = |\boldsymbol{x}_{measured} - \boldsymbol{x}_{ideal}|$. From the analysis, we estimate the input error follows a Gaussian distribution ~N(0.002, 0.0008).

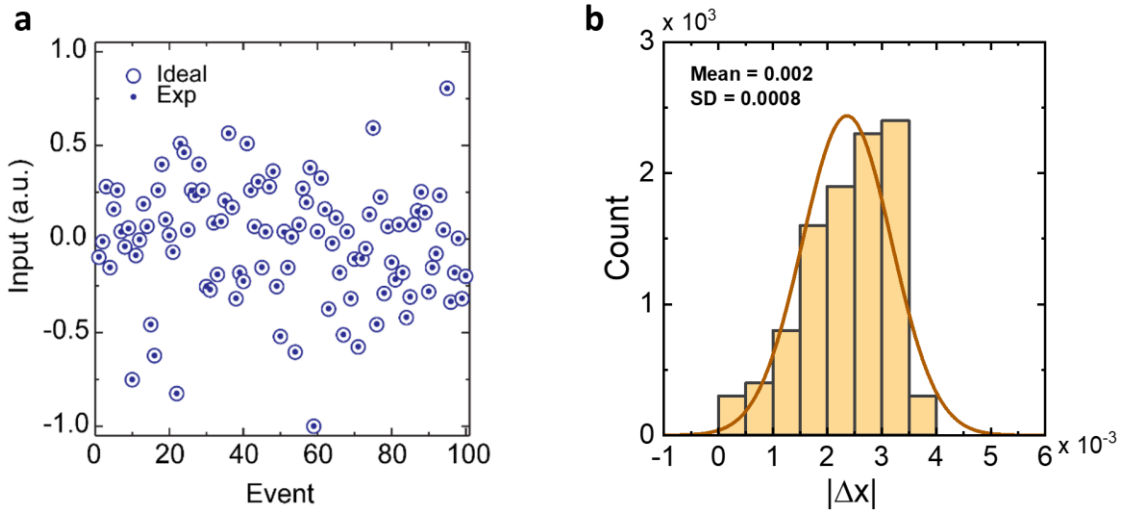

Figure 3.4 Input error characterization. **a.** An encoded optical signal trace is used as the input in the first layer obtained from measurement with its corresponding ideal value. The circles and the solid dots are the ideal value and the corresponding measured value respectively. **b.** The histogram of the input signal error shows that the SD of the input signal error is negligible, only 8×10^{-4}.

3.3.3 *Characterization of weight setting error (write error)*

Figure 3.5 shows the evolution of Γ during the programming process of using optical control pulses to set negative (-0.7), zero (0.0), and positive (0.7) values, respectively. We implement the network model on a 2×2 tensor core with four PMMCs (Figure 3.1d). The kernel weight W_{ij}^{l} value is mapped to the corresponding mode contrast Γ_{ij}^{l} as $\Gamma_{ij}^{l} = W_{ij}^{l} \cdot \left(|\Gamma|_{max}^{l} / |W|_{max}^{l} \right)$, where $|\Gamma|_{max}^{l}$ is the maximum absolute mode contrast, $|W|_{max}^{l}$ is the maximum absolute kernel weight of layer l. Given the limited number of PMMCs on a chip, we repeatedly reset the kernel elements on the same devices, which bottlenecks the computing speed. With a sufficiently large tensor core in a photonic crossbar array architecture, one could directly map the full kernel matrices to the hardware so the computing speed will be much accelerated.

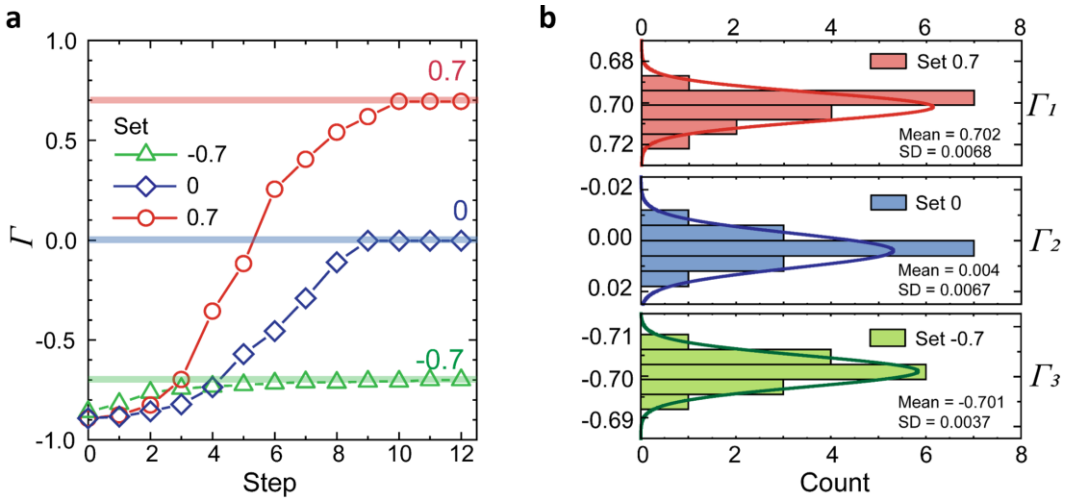

Figure 3.5 Weight write error characterization. **a.** Process of programming the mode contrast of a kernel element using optical pulses. The target Γ values are -0.7, 0, and 0.7, respectively. **b.** Histogram of Γ value distribution when the kernels are repeatedly set to be -0.7, 0, and 0.7, respectively. The SD for each set is 0.37%, 0.67%, and 0.68%, respectively.

The analog nature of weight programming and data encoding and transferring in the photonic neural network limits the precisions of MVM calculations and makes the computation error-prone. The computation errors would accumulate through the layers of the network and impair the final results. To quantify the kernel weight setting errors (write errors) in our system, we repeatedly program different fixed Γ values and estimate the short-term inaccuracy by measuring the variation $\Delta\Gamma$. Figure 3.5b shows that the SD of 15 programming operations is less than 0.7%, which is one order of magnitude larger than the input encoding error. Thus, for the short-term error, the programming inaccuracy $\Delta\Gamma$ (write error), limited by the inaccuracy of the programming optical pulses, is one of the dominant error sources. Figure 3.6 shows the corresponding SD of kernel weight with the various numbers of bits of resolution. The 0.7% write error limits the achievable resolution to 6 bits of the resolution.

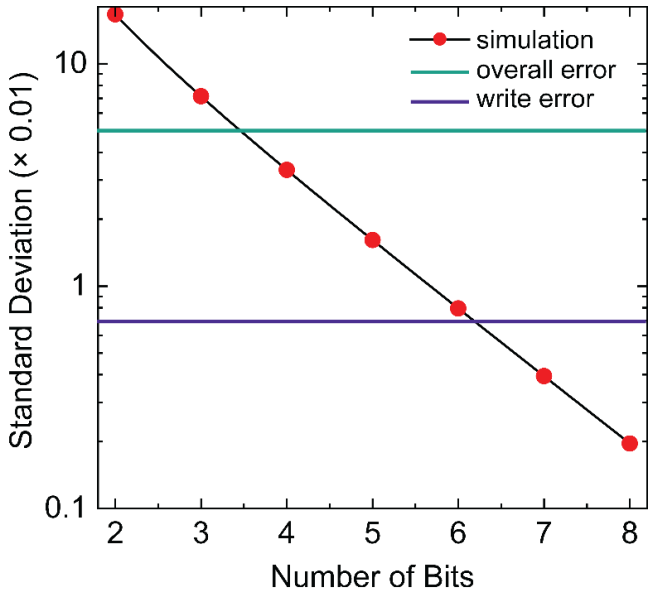

Figure 3.6 The corresponding SD of kernel weight with the various numbers of bits of resolution. The purple line and green line indicate the SD of the photonic kernel weights considering the write error (purple line) and the total error (green line), respectively. The write error corresponds to 6 bits of the resolution, and the overall error corresponds to over 3 bits of the resolution.

3.3.4 *Characterization of long-term read error*

Another error source is the long-term measurement fluctuations (read error), including the noise of photodetectors, the variation of the O/E and E/O conversions, and the thermo-optic fluctuation of the PCM. These errors collectively contribute to an effective error $\Delta W_{ij}^{l} = \left(\left| \Gamma \right|_{max}^{l} / \left| W \right|_{max}^{l} \right) \cdot \Delta \Gamma_{ij}^{l}$ on the kernel element weight W_{ij}^{l}, where $\Delta \Gamma_{ij}^{l}$ is the total write error. To estimate the computation error of the overall system, Figure 3.7 compares the measured MVM error distributions with the simulation, which assumes a Gaussian distribution of error with zero mean. The MVM result estimates the overall error $\Delta \Gamma_{ij}^{l}$ to be 5%, corresponding to over 3 bits in resolution, which we subsequently use in the noise-aware training and simulation later in this chapter.

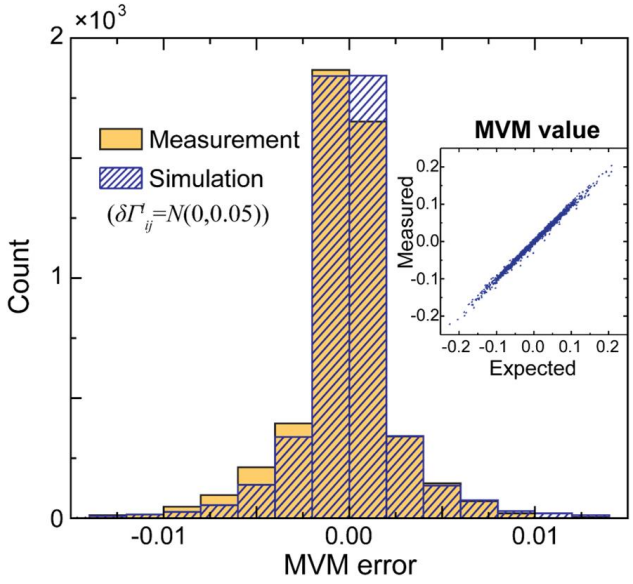

Figure 3.7 Long-term read error characterization. Histograms of the error distribution in the experimental measurement (solid) and the simulation (hashed) when assuming the $\Delta\Gamma^{l}_{ij}$ follow a Gaussian distribution with an SD of 5%. Inset: Measured MVM accuracy for 4900 MVM operations in the first layer of the network.

3.3.5 *Compare the error levels in the MVM operation using the PMMC-based photonic tensor core*

The ratio between the two error terms that contribute to the element y_i^l of layer *l* during the MVM calculation is estimated by:

$$\frac{\left(\Delta \mathbf{W}^{\boldsymbol{l}} \cdot \boldsymbol{x}^{\boldsymbol{l}}\right)_{\boldsymbol{i}}}{\left(\mathbf{W}^{\boldsymbol{l}} \cdot \Delta \boldsymbol{x}^{\boldsymbol{l}}\right)_{\boldsymbol{i}}} = \frac{\sum \Delta w^l{}_{ij} x_j^l}{\sum w^l{}_{ij} \Delta x_j^l} \approx \frac{\dfrac{\left|\Delta w^l\right|}{\left|w^l\right|_{ave}}}{\dfrac{\left|\Delta x^l\right|}{\left|x^l\right|_{ave}}} \approx \frac{\dfrac{\Delta \Gamma^l}{\left|\Gamma^l\right|}}{\dfrac{\delta\left|x^l\right|}{\left|x^l\right|}} \tag{3.6}$$

where $\Delta\Gamma^l$ and $\delta|x^l|$ are the standard deviations of the mode contrast setting error and the input error, $\langle|\Gamma^l|\rangle$ and $\langle|x^l|\rangle$ are the mean values of the mode contrast setting error and the input error. Take the first layer as an example, the $\langle|x^1|\rangle$ and $\delta|x^1|$ are 0.1723 and 8×10^{-4}, respectively (see Figure 3.4). The $\langle|\Gamma^1|\rangle$ and $\Delta\Gamma^1$ are 0.4236 and 0.05 for weight setting error (0.007 for short-term write noise $\delta\Gamma$). The ratio between the two error terms is 25.42 (4.07 only short-term write noise considered), thus the input encoding error is not the main error source in our photonic GAN while the write error and the long-term read error are the dominant error sources.

3.4 The architecture of the GAN generates the handwritten number

3.4.1 *GAN architecture used to generate handwritten numbers*

Figure 3.8 plots the architectures of the network models used later in this Chapter. The generator models in the GAN (see Figure 3.8a and Figure 3.8c) consist of a fully connected layer (FC) and deconvolution layers (Deconv). Between each hidden layer, batch normalization is operated before applying the nonlinear function. For the generator to generate “7” (Figure 3.8a), we choose the LeakyReLu as the nonlinear function after each hidden layer. For the generator to

generate full digits (Figure 3.8c), we choose the ReLu as the nonlinear function. Both models take four random inputs to generate a single image. We use the hyperbolic tangent function as the nonlinear function at the final output. Only the generator used to generate the number "7" is realized by the photonic hardware experimentally. The generator (Figure 3.8c) used to generate ten digits is realized on a digital computer due to the experimental limitation of the photonic kernel size. Table 3.1 and Table 3.2 list the data flow of the two generators and the total number of parameters in each layer. Specifically, in the number "7" generation experiment, all the FC layers and the Deconv layers are performed using photonic hardware with 325 parameters in total stored in the photonic tensor core (See Table 3.1). The batch normalization and nonlinear function are performed using a computer.

The discriminators in the GAN (see Figure 3.8b and Figure 3.8d) consist of an FC layer and several convolution layers (Conv). We choose the LeakyReLu as the nonlinear function after each hidden layer and apply the sigmoid function at the final output for both discriminator models. Both discriminators are trained on a digital computer.

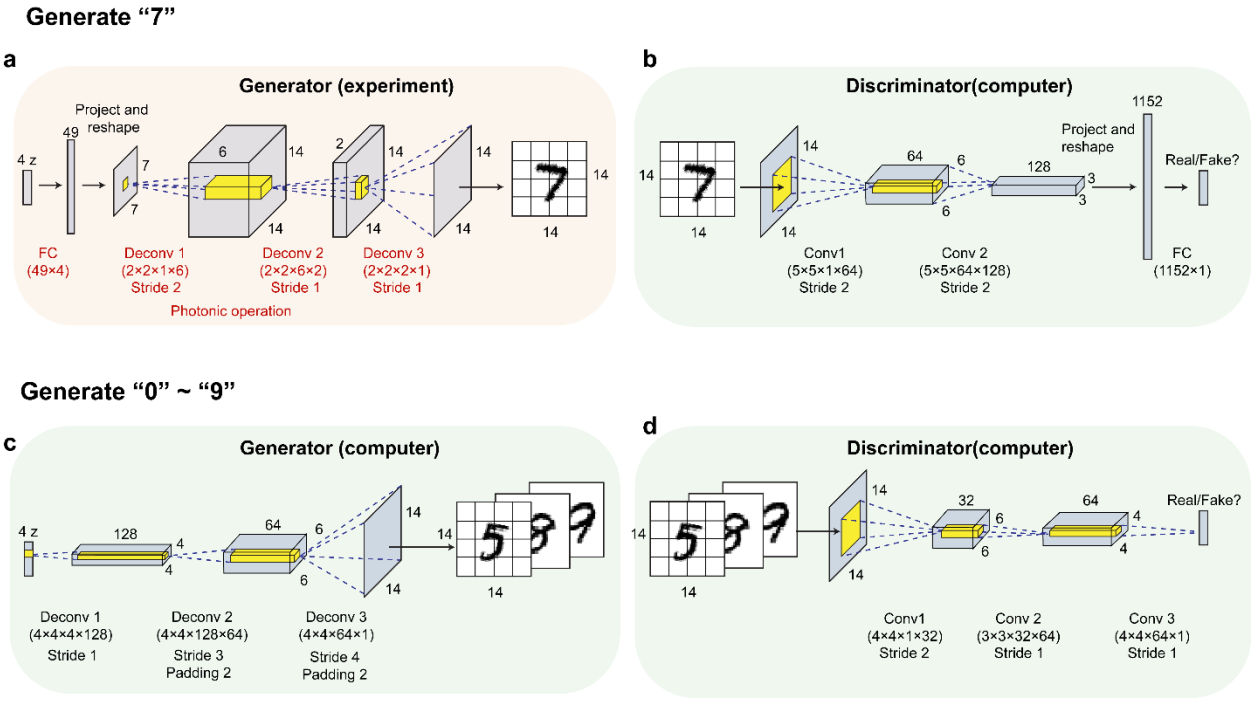

Figure 3.8 Experimental and simulational generator architecture. **a** and **b.** The architecture of the generator (**a**) and discriminator (**b**) in GANs are used to generate the handwritten number "7". The generator is experimentally demonstrated on a photonic tensor core (shaded red) while the discriminator is realized on a digital computer (shaded green). In the generator, the fully connected (FC) and deconvolutional (Deconv) layers are implemented with the photonic tensor core. The post-processing steps, including the batched normalization and nonlinear functions, are achieved with electronic hardware. **c** and **d.** The architecture of the generator (**c**) and discriminator (**d**) in GANs that are used to generate the ten handwritten digits from "0" to "9". Both the generator and discriminator are trained on a digital computer.

Table 3.1 The data flow in the generator generates the number "7". The operation steps, input/output matrix dimensions, kernel sizes, the total number of parameters, and the implementation method ("O" for optically and "E" for electrically) of each layer are listed. In this generator, the FC layer and three Deconv layers are all performed experimentally using the photonic hardware with 325 parameters stored in the photonic tensor core.

#Layer	Operation	Input dimensions	Kernel dimensions	Output dimensions	# of parameters	Method (O/E)
1	input	/	/	4×1×1	0	/
2	FC	4×1×1	49×4×1×1	49×1×1	245	O
3	batch	49×1×1	/	49×1×1	196	E
4	leaky ReLu	49×1×1	/	49×1×1	0	E
5	reshape	49×1×1	/	7×7×1	0	E
6	Deconv 1	7×7×1	2×2×1×6	14×14×6	24	O
7	batch	14×14×6	/	14×14×6	24	E
8	leaky ReLu	14×14×6	/	14×14×6	0	E
9	Deconv 2	14×14×6	2×2×6×2	14×14×2	48	O
10	batch	14×14×2	/	14×14×2	8	E
11	leaky ReLu	14×14×2	/	14×14×2	0	E
12	Deconv 3	14×14×2	2×2×2×1	14×14×1	8	O
13	output (tanh & normalize)	14×14×1	/	14×14×1	0	E

Table 3.2 The data flow in the generator to generate the number “0”~”9”. The operation steps, input/output matrix dimensions, kernel sizes, and total number of parameters of each layer are listed. This generator is performed on a digital computer.

#Layer	Operation	Input dimensions	Kernel dimensions	Output dimensions	# of parameters
1	input	/	/	4×1×1	0
2	Deconv 1	4×1×1	4×4×4×128	4×4×128	8192
3	batch	4×4×128	/	4×4×128	512
4	ReLu	4×4×128	/	4×4×128	0
5	Deconv 2	4×4×128	4×4×128×64	6×6×64	131072
6	batch	6×6×64	/	6×6×64	256
7	ReLu	6×6×64	/	6×6×64	0
8	Deconv 3	6×6×64	4×4×64×1	14×14×1	1024
9	output (tanh & normalize)	14×14×1	/	14×14×1	0

3.4.2 *Estimation of the diversity of the generated images using FID as a criteria*

The Frechet inception distance (FID) is a metric for evaluating the quality of generated images in both fidelity and diversity. It was proposed specifically to evaluate the performance of generative adversarial networks [98]. To calculate FID, we design another CNN, as shown in Figure 3.9, to classify the generated handwritten digits. The overall error rate of this CNN is only 2.43% after training. When a 14×14 image is sent into this network, the activation features obtained before the last FC layer is reshaped as the “feature vector” with the size of 160 ×1, and the *n* feature vectors obtained after *n* images are fed into the CNN are combined to form a feature matrix with the size of 160 × *n*. The FID evaluates the generated images by statistically comparing

them with the real images from the target domain. Assuming the matrice **X** and **Y** are the feature matrices of the GAN-generated images and the real images from the MNIST database, respectively, the FID is defined as:

$$FID = \mu_X - \mu_Y{}^2 + tr\left(\Sigma_X + \Sigma_Y - 2\left(\Sigma_X\Sigma_Y\right)^{\frac{1}{2}}\right) \tag{3.7}$$

where μ_X and μ_Y refer to the feature-wise mean vectors of the GAN-generated images and real images, Σ_X and Σ_Y are the covariance matrices of the corresponding feature matrices **X** and **Y,** and "*tr"* is trace operation. To increase FID accuracy, we also break the generated images into groups and calculate the mean and SD of the FID. **We highlighted here that the lower the FID score, the better the performance of the GAN.**

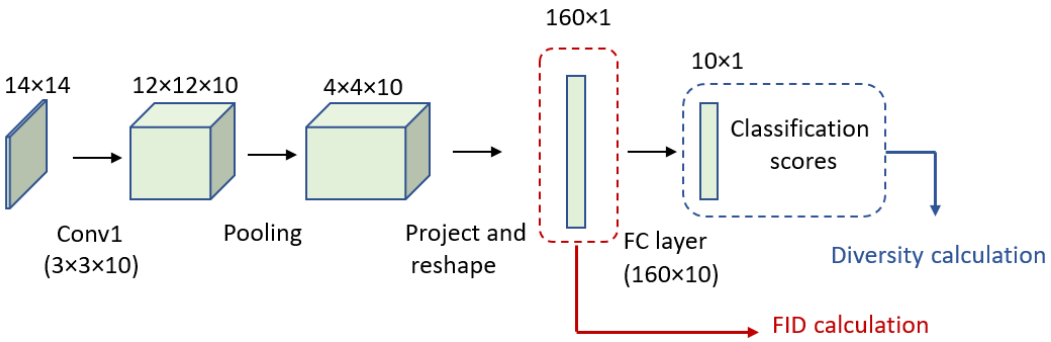

Figure 3.9 The architecture of the convolutional neural network used for handwritten digits classification. When a 14×14 image is sent into this network, the activation features obtained before the last FC layer are reshaped as the "feature vector" with the size of 160 ×1 to calculate the FID. The statistics of the classification results are used to estimate the diversity.

We also estimate the diversity of generated images using the diversity coefficient, which is defined as the SD of the percentage of each number class in the generated images:

$$Diversity\ Coefficient = \sqrt{\frac{\sum_{i=0}^{9}\left|P_i - \frac{1}{10}\right|^2}{10}} \tag{3.8}$$

where P_i is the statistic percentage for the generated image that is successfully classified as the number "*i*" (0 to 9). **The lower the diversity coefficient, the higher the diversity of the generated images with the more uniform distribution in the ten classes** [101]. For example, the MNIST training database has a close-to-uniform percent of ten classes, so its diversity coefficient is as low as 0.00585.

3.5 Noise-aware training of the photonic generative model

3.5.1 *Noise effect on the GAN*

Unlike the discriminative network, where the input regularities or patterns are well-defined, the generator network takes random numbers as the input. It would be more susceptible to the effective weight setting noise ΔW_{ij}^{l}, which could degrade the quality of the generated new instances [102]. To reveal the noise effect on the GAN, we emulate the noisy hardware on a GAN model that is trained using a noiseless offline training approach but add a random error ΔW^{l}_{ij} (introduced by $\Delta \Gamma^{l}_{ij}$ with a Gaussian distribution $N(0, 0.05)$) when using it to generate images. Figure 3.10 plots 100 images of 14×14 pixels generated from simulation using random inputs produced by the optical RNG. These images show the handwritten "7" but with very noisy backgrounds, demonstrating that the noise-free training algorithm is impaired by the practical weight setting noise.

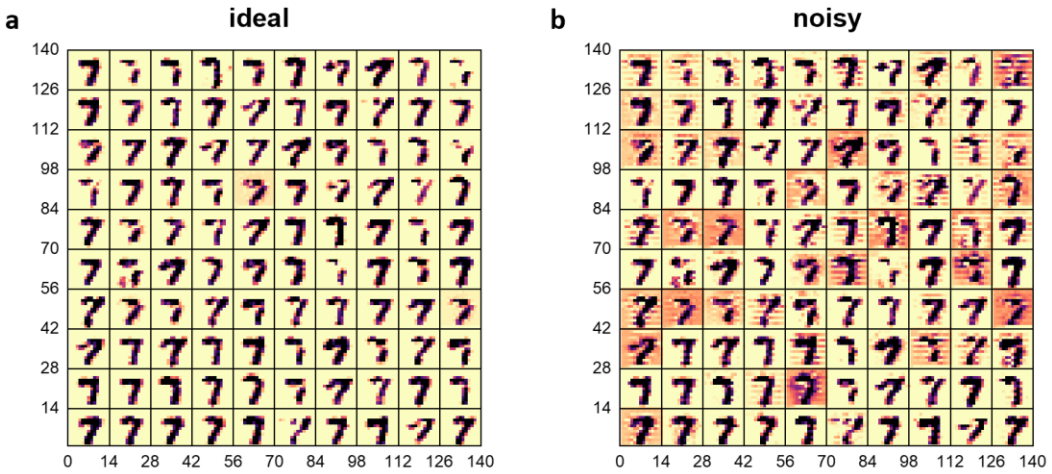

Figure 3.10 100 images (size: 14×14 pixels) generated by noise-free GAN and the noisy GAN under effective kernel weight setting error (introduced by 5% Gaussian random error $\Delta\Gamma^{1}_{ij}$) and using random inputs~ $N(0,0.2)$ produced by the optical RNG. All images are generated from simulation results.

Therefore, it is necessary to consider hardware noise during training to realize a GAN that is resilient to realistic noises. Theoretically, it has been proven that adding noises to the training data of a neural network is equivalent to an extra regularization added to the error function, which can significantly improve hardware noise tolerance in a discriminative neural network. Meanwhile, it was shown that introducing noise on kernel weights during training enhances the robustness against weight perturbations of multi-layer perceptrons, such that inference accuracy close to the software baseline could be achieved. However, previous demonstrations of noise-aware solutions are limited to discriminative networks. In the following sections, we will extend the theoretical, simulation, and experimental validations of effective noise-aware solutions to photonic GAN networks.

3.5.2 *Noise-aware training of the photonic generative model*

For our photonic GAN, we propose and experimentally validate three noise-aware training approaches, namely, the input-compensatory approach (IC-GAN), the kernel weight-compensatory approach (WC-GAN), and the curvature regularization approach (CR-GAN) to improve the tolerance of the network to the effective weight setting noise ΔW^l_{ij}. The IC-GAN approach inflates the SD of the random signal input from the experimental value of 0.2 V to 0.5 V during training. The WC-GAN approach adds ΔI^l_{ij} with 5% STD to the corresponding weight at each forward-propagation pass but performs noiseless gradient descent in the back-propagation pass. The CR-GAN evolves from the WC-GAN and is used to improve the GAN robustness further by adding a curvature regularization term in the loss function.

Figure 3.11 shows the three noise-aware training approaches we proposed in the main text, the IC-GAN, the WC-GAN, and the CR-GAN. All three approaches are based on the offline training configuration that the GAN is first trained on a digital computer. After training, we mapped the obtained parameters to the photonic hardware for experimental implementation.

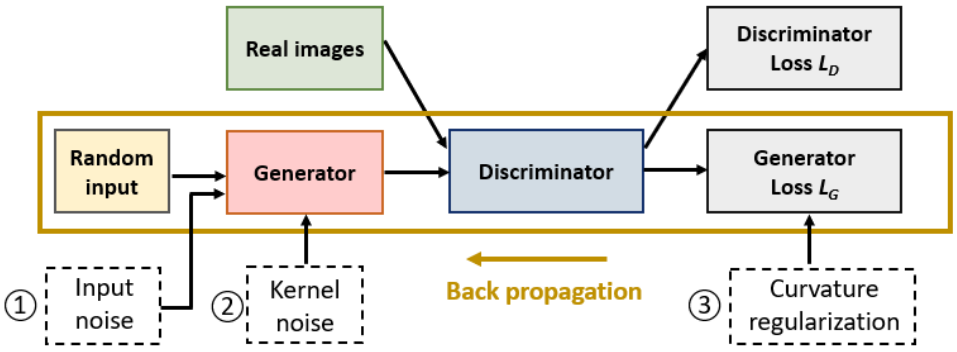

Figure 3.11 Schematic diagram of the noise-aware training approaches of the GAN. The IC-GAN approach inflates the STD of the random input (labeled as process ①) and keeps all the other steps the same as the conventional NF-GAN training algorithm. The WC-GAN performs noiseless gradient descent and weight update in the back-propagation process but adds additional noise onto the corresponding kernel weight (labeled as process ②) in each forward-propagation process. For the CR-GAN, a curvature regularization term is added to the total loss function, $L_{wr}=L_G+rL_r$, during each forward-propagation process (labeled as process ③) before performing noiseless gradient descent and weight update in the back-propagation process.

The IC-GAN approach inflates the STD of the random signal input, from 0.2 to 0.5 in our case, during training of the network, while all the other steps in the forward and backward-propagation pass such as the loss function calculation, the gradient descent, and weight update, are the same as the conventional GAN training algorithm. We control the learning rates for the discriminator and the generator at the same level of 1×10^{-4} to avoid one overpowering the other. Once all the parameters are obtained, we implement the network experimentally where the input noise has a smaller SD (0.2 in our experiment).

For the WC-GAN approach, we cast the effect of all the error sources into the 5% SD Gaussian noise $\Delta\Gamma^{l}_{ij}$ and add additional noise onto the corresponding kernel weight at each forward-propagation pass following the equation, $W^{l'}_{ij}= \Delta W^{l}_{ij}+ W^{l}_{ij}$, where the ΔW^{l}_{ij} is given by $\Delta W^{l}_{ij} = \frac{|\boldsymbol{W}|^{l}_{max}}{|\boldsymbol{\Gamma}|^{l}_{max}} \cdot \Delta\Gamma^{l}_{ij}$. As shown in Figure 3.12a, the updated kernel weight causes a deviation of the loss function $L^{'}_{G}$ from its ideal value L_G, thus leading to a different gradient. The WC-GAN performs noiseless gradient descent and weight update in the back-propagation pass based on the $L^{'}_{G}$. After every update of the kernel weight, the absolute value of the maximum weight $|W|^{l'}_{max}$ may expand, while the total noise injected ΔW^{l}_{ij} on the corresponding kernel weight during training is proportional to the absolute value of the maximum weight, the noise may grow uncontrollably with growing maximum weight values and prevent the training from converging. Therefore, we clip the kernel weight distribution in the desired range by rescaling the kernel element with the maximum absolute value of each layer by a factor (0.995 in training) after every weight update. The weight clipping process improves the GAN training convergence. All the other steps in the WC-GAN training approach are the same as the conventional algorithm.

For the CR-GAN approach (see Figure 3.11 and Figure 3.12b), we hypothesize that the robustness against experimental noise can be maximized if the distance between the weight gradient at the point $\mathbf{W}^{l}$ in parametric hyperspace and the gradient of neighborhood points $\mathbf{W}^{l'}$ are minimized. We define the regularization term L_r as the maximal distance between the weight gradient and the gradient of the neighborhood point $L_r = \frac{dL_G}{d\mathbf{W}^{l}} - \frac{dL^{'}_{G}}{d\mathbf{W}^{l'}}$, where the range of the neighborhood is defined by the discretization step h. In each training step, the regularization term is added to the total loss $L_{wr}=L_G+rL_r$ where r is the regularization strength, and then the back-

propagation is performed on the weights. In our simulation, *h* and *r* are varied under various noise levels to obtain optimal performance.

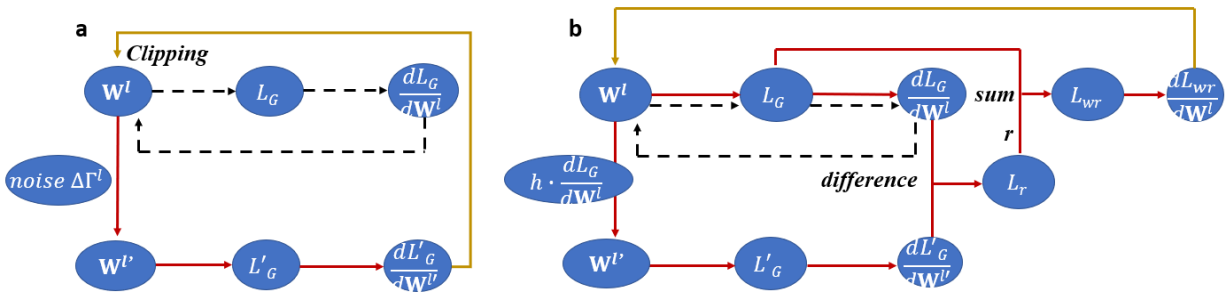

Figure 3.12 Schematic of the kernel weight update process flow for WC-GAN (**a**) and CR-GAN (**b**), respectively. The solid red line and the solid yellow line indicate the forward-propagation pass and the backward-propagation pass, respectively. The dashed black line indicates the conventional GAN training process.

3.5.3 *Performance comparison between noise-aware training approaches*

Figure 3.13b and Figure 3.13c show the experimentally generated images of handwritten "7" by the photonic GAN trained using both approaches. For a fair comparison, the random number inputs used for inferences are produced by the same optical RNG. Compared to the images generated by the noise-free trained GAN (NF-GAN) (Figure 3.13a), the images generated using both noise-aware approaches display much clearer patterns with lower background noise, thus validating the noise tolerance of the IC-GAN and WC-GAN. Furthermore, we observe that the images generated by the WC-GAN (Figure 3.13c) have richer handwritten-like features than those by the IC-GAN (Figure 3.13b), with more diverse variations in styles. Therefore, we conclude that the WC-GAN is advantageous for practical implementation using non-ideal analog hardware.

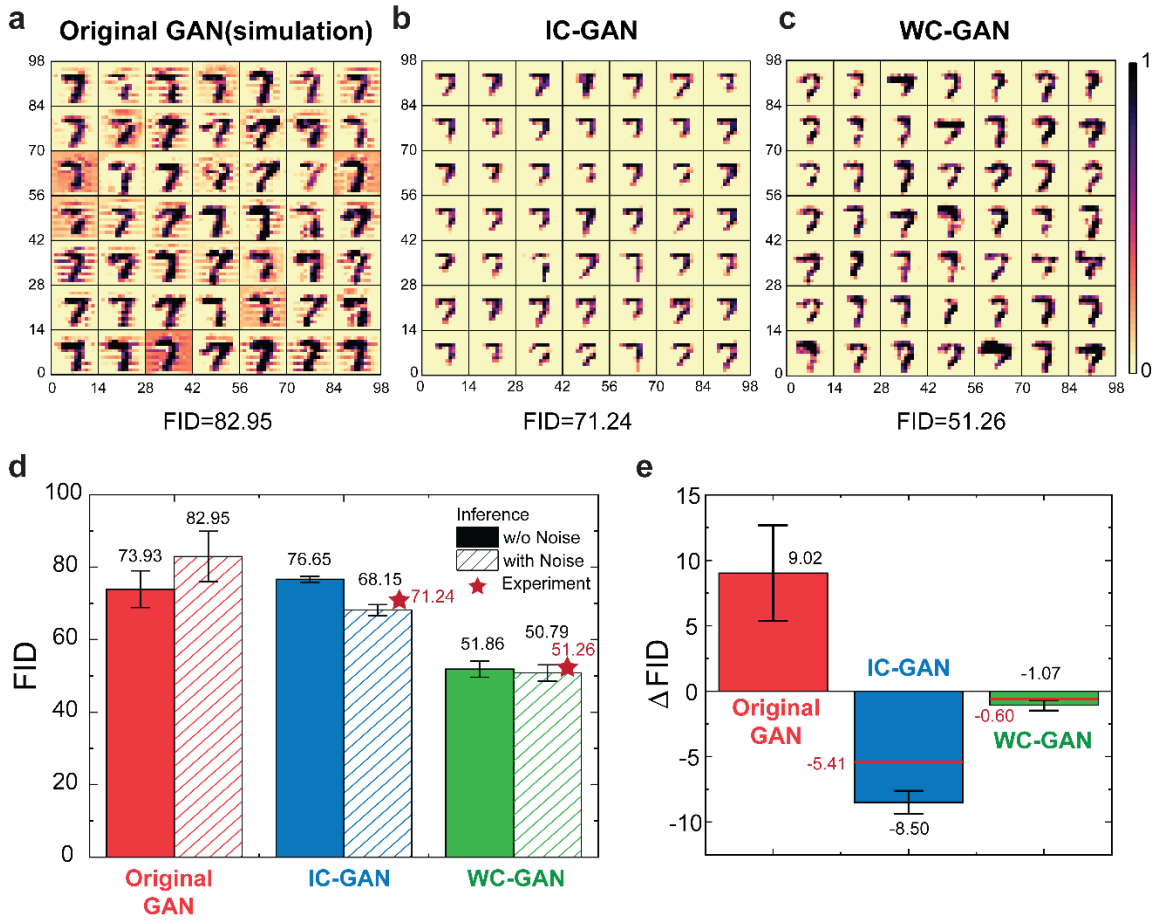

Figure 3.13 Generating handwritten numbers with GAN. **a-c**. 49 images (size: 14×14 pixels) generated by NF-GAN (**a**), IC-GAN (**b**), and WC-GAN (**c**) under effective kernel weight setting error (introduced by 5% Gaussian random error $\Delta\Gamma^{1}_{ij}$) and using random inputs~ $N(0,0.2)$ produced by the optical RNG. **a** is generated by simulation, **b** and **c.** are from the experiments. **d.** The FIDs of the generated images, assuming the network is trained using various approaches and is implemented either on the ideal (solid bars) or noisy hardware (hashed bars). The FIDs obtained from the experimental results are labeled as stars. **e.** The difference of FID (ΔFID) in (**d**). The ΔFIDs from the experimentally generated images are denoted by the red lines.

To quantitatively compare the GAN performance, we use the standard FID [101], which evaluates both the fidelity and diversity of the generated images by comparing the feature distribution in the generated images with images from the training dataset. Recall that the lower the FID score, the better the performance of the GAN. In Figure 3.13d, the FIDs of the images generated by the NF-GAN, the IC-GAN, and the WC-GAN, respectively, are compared, assuming either ideal (FID_{ideal}) or noisy (FID_{noisy}) hardware. The FID_{noisy} (hashed bars in Figure 3.13d) is

the lowest for the WC-GAN and the highest for the NF-GAN, consistent with the observation in Figure 3.13a-c. The impact of hardware noise $\Delta FID = FID_{noisy} - FID_{ideal}$ is plotted in Figure 3.13e. The noise-aware WC-GAN and IC-GAN show two notable benefits. First, the FID_{ideal} (solid bars in Figure 3.13d) for the WC-GAN is lower than the NF-GAN (e.g., the software baseline), indicating that introducing noises during training helps GAN learn better. Such a gain is absent in discriminative networks, where the inference accuracies of the noise-aware trained model cannot exceed the software baseline. Second, surprisingly, the noise impact results (Figure 3.13e) show that, unlike the NF-GAN, the WC-GAN and IC-GAN implemented on the photonic hardware with practical noise (hashed bars in Figure 3.13d) perform even better in inference than the noiseless hardware (solid bars in Figure 3.13d). In contrast, a discriminative network's inference accuracy always decreases with more noisy hardware. This surprising gain in performance suggests photonic neural networks' potential in generative models despite the inevitable optoelectronic noises and errors.

To predict if the noise-aware approach performance gain is scalable, in simulation, we train a larger-scaled GAN to generate images of all 10 number digits, using ideal or noise-aware approaches under various levels of writing errors. Figure 3.14a shows the FID score of the results as a function of ΔT^{1}_{ij}. Here, CR-GAN is used to improve the GAN robustness further (see Section 3.5.2 for more details about the CR-GAN). The comparison shows that the CR-GAN performs better than the NF-GAN at every error level. Note that under our present realistic noise level of 5% (Figure 3.13e), the FID of CR-GAN is still below the software baseline, whereas the NF-GAN's FID is higher than the baseline. For both approaches, with the increasing noise level, the FID first drops until reaching a minimum at ~2.5% noise and then increases. To explain this, we further examine the images generated by CR-GAN at three noise levels: 0%, 5%, and 10% in Figure 3.14b-d.

The comparison shows that the increasing hardware noise in GAN would improve the diversity but at the same time reduce the fidelity of the generated images. The trade-off results in a minimal FID at ~2.5% noise, as shown in Figure 3.14a. Throughout the full range of noise levels, the noise-aware approach consistently improves the GAN over the noiseless approach.

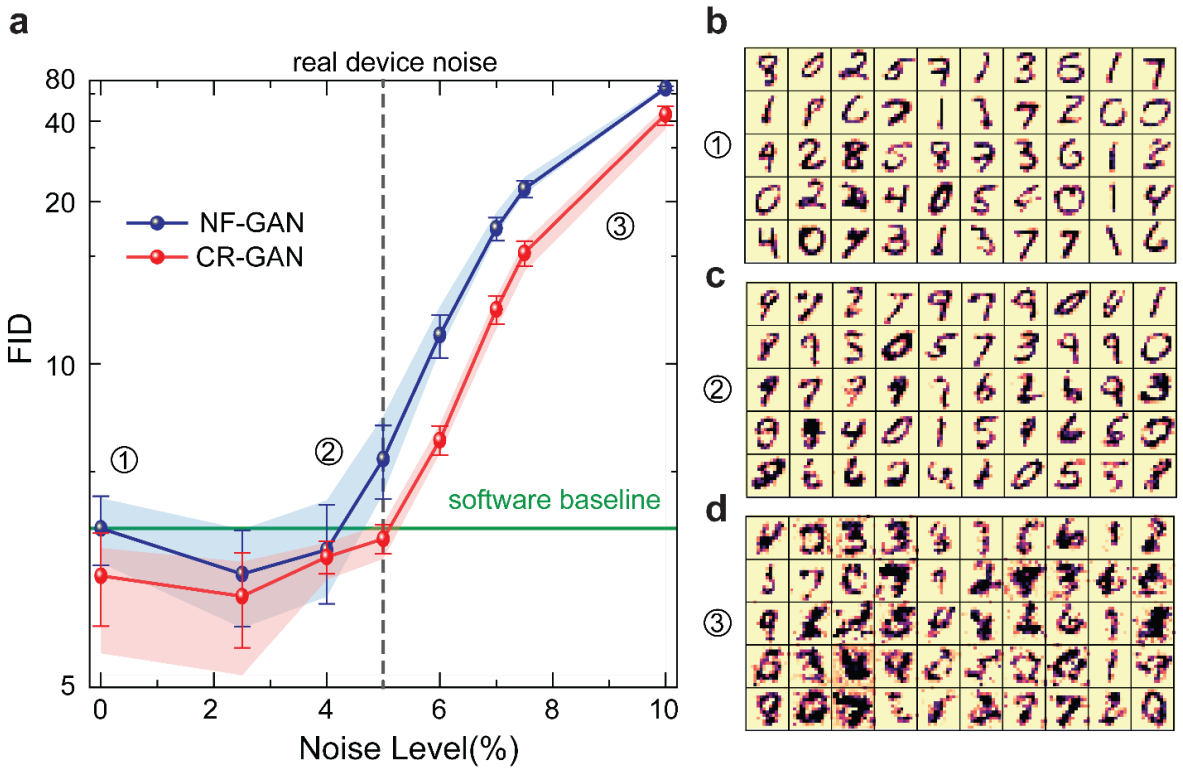

Figure 3.14 Scalability of noise-aware training. **a.** The FID of the generated images by the NF-GAN and the CR-GAN, respectively, under various effective mode contrast setting noise $\Delta\Gamma^{1}_{ij}$ with SD ranging from 0% to 10%. The shaded region indicates the range of FID over 5 individual tests. The FID is lower for CR-GAN at every noise level. At the measured noise level of 5% (black dashed line), the FID for CR-GAN is below the software baseline (solid green line) while the FID for the NF-GAN is above it. **b-d**. 50 images (size: 14×14) generated by CR-GAN assuming effective mode contrast setting noise of 0% (**b**), 5% (**c**), and 10% (**d**).

3.6 Conclusion and outlook

In conclusion, in this chapter, we demonstrate a photonic generative network based on phase-change photonics, which is used to form a GAN network and utilize the intrinsic noise sources in the photonic system. Unlike the previously demonstrated discriminative networks that suffer from hardware noise and errors, our experimental and simulation results show that the photonic generative network can not only tolerate but also benefit from a certain level of hardware noise after training by noise-aware training approaches. Our finding expands the current implementation

of photonic neural networks to generative models [103], in which the inevitable and ubiquitous optoelectronic noises and errors can be mitigated and even leveraged in intelligent ways. We emphasize that the proposed noise-aware training approaches are generic and thus applicable to various types of optoelectronic neuromorphic computing hardware. The improved noise resilience of the model also implies their scalability in large-scale photonic neural networks with tightly co-integrated electronics and photonics.

Chapter 4. Freeform Direct-write and Rewritable Photonic Integrated Circuits in Phase-Change Thin Films

The application of photonic integrated circuits (PICs) is rapidly spreading in diverse technology domains, ranging from computing [104–106], communication [4,107], sensing [108,109], and quantum technology [12,15]. Traditionally, PICs are fabricated in thin film materials, including silicon [57], silicon nitride [110], indium phosphide [111,112], and lithium niobates [30,31], using nanofabrication processes including lithography, etching, and deposition on tools installed in cleanroom facilities. Compared with electronics, where prototypes can be rapidly built by students using discrete elements plugged into a breadboard, photonics research faces the barriers of the limited accessibility to fabrication facilities, the high costs associated with the fabrication process, and the extended design-to-device turnaround time. The combination of these barriers impedes widespread innovation and throttles the broader impact of photonics research and education.

Moreover, there is a growing interest in programmable PICs to realize highly flexible photonic networks for emerging applications, including optical computing [106], optical interconnect for neural network accelerators [4] and quantum computing [12,13,113]. Currently, programmable PICs are realized through a network of tunable components, including phase shifters, directional couplers, and interferometers, connected in a highly complex architecture [1,107,114]. The programmability of those schemes has been limited to switching and rerouting the network, while freeform reconfiguration of the system's functionality remains difficult to achieve. As the complexity increases, programmable PICs face overwhelming challenges of scalability, programming precision, and flexibility, which limit their practicality and increase cost [115–117]. Another challenge PICs face is the waste generated from malfunctioning dies and chips during the prototyping and testing stage, which are virtually impossible to repurpose once completed – a challenge that could be overcome using rewritable techniques.

In this chapter, we report a simple, fast turnaround and low-cost approach to creating and reprogramming PICs that could shift the paradigm in photonics research, prototyping, and education [118]. Our technique is based on direct laser writing (DLW) on phase-change material (PCM) thin films. The method writes the PICs in only one optical patterning step and without using any traditional lithography and etching processes. The PICs are created by utilizing PCM's

dramatic refractive index contrast between the two nonvolatile phases, amorphous and crystalline, which are reversibly switchable using optical pulses. Previously, laser writing to control phase transition in PCMs has been used in optical storage, such as rewritable compact disks (CD-RW) [42]. Free-space optical switching of PCMs has also been used to realize reconfigurable metasurface [119,120], rewritable color displays [121], and polariton nanophotonic devices [122]. Laser patterning of photonic circuits has been reported, including femtosecond laser writing of waveguides in silica [123–126], chalcogenide glass [127–130] and a programmable III–V semiconductor-based photonic device [131]. However, laser writing of glass waveguides is a non-reversible process, whereas the optical patterning in III-V semiconductors is volatile, requiring continuous laser irradiation to sustain. Additionally, none of the above techniques creates complete end-to-end PIC systems with multiple reconfigurable functions, as we report in this chapter.

By eliminating the reliance on traditional fabrication processes, our technique enables researchers to explore a wider space of design possibilities and system functionalities more rapidly.

Moreover, such a technique will allow researchers and students who do not have access to nanofabrication facilities to prototype and reuse PIC designs and thus democratize photonics research to a broader community. It will enable more students and educators to engage in hands-on experimentation, thereby fostering innovation and knowledge dissemination and generating a broader impact to promote workforce development in photonics.

4.1 Concept of direct-write and rewritable PICs

Figure 4.1a illustrates the concept of direct writing, erasing, and rewriting PICs. The PIC is written on a standard oxidized silicon substrate, which is coated with a 200-nm-thick SiO_2 layer covering a 30-nm-thick Sb_2Se_3 layer on a 330-nm-thick Si_3N_4 film. The SiO_2 capping layer protects and prevents oxidation of the Sb_2Se_3 layer. Waveguiding in the Sb_2Se_3 film is achieved by using the crystalline phase (cSb_2Se_3) as the high-index core and the amorphous phase (aSb_2Se_3) as the cladding (Figure 4.1b). This binary phase configuration, that is, no mixed phases, enables

the confinement of the fundamental transverse electric (TE_0) optical mode within a cSb_2Se_3 waveguide, assisted by the underlayer of Si_3N_4, as shown in the simulated mode profile in Figure 4.1c.

We directly write the circuit layout on a blank cSb_2Se_3 thin film using a commercial laser writing system (Heidelberg DWL 66+, 405 nm laser), mainly to leverage its precision stage and computer control system for large area writing. Alternatively, a homebuilt system using an off-the-shelf laser diode is sufficient to write a smaller area. To write the PICs, the focused laser beam with a power of 27.5 mW is scanned across the film at a speed of 3.0 mm^2/min, inducing a controlled phase transition from cSb_2Se_3 to aSb_2Se_3 to create the claddings and define the core. This laser-controlled phase transition creates a circuit with a resolution limited by optical diffraction. Figure 4.1d shows a series of rectangular aSb_2Se_3 structures created by DLW with varying widths ranging from 1 μm to 200 nm. The minimum achievable feature size is 300 nm, exceeding the resolution reported in previous works [46,120,121]. Moreover, the circuit layout can be erased either locally or globally. Local erasure is achieved by scanning the laser beam at a lower speed (0.1 mm^2/min) with a reduced laser power of 15 mW, inducing a phase transition from aSb_2Se_3 back to cSb_2Se_3 in desired areas (Figure 4.1e). This capability enables modifications and corrections to the PIC design. Alternatively, global erasure can be accomplished by heating the whole substrate to above 180 ℃, promoting a complete phase transition across the entire PCM film. Thereby, the DLW technique provides very versatile capabilities for writing and rewriting phase-change PICs in freeform without using any advanced fabrication tools.

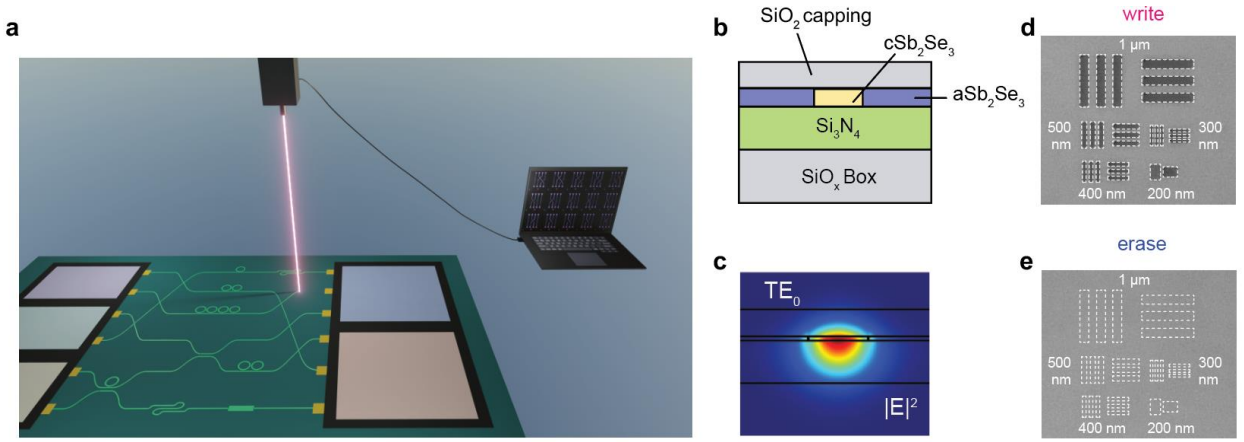

Figure 4.1 Direct-write and rewritable phase-change photonic integrated circuits. **a**. Artistic illustration of freeform writing and rewriting PICs on Sb_2Se_3 thin film. **b**. The cross-sectional view of a cSb_2Se_3 optical waveguide structure. The waveguide is directly written in the PCM thin film without fabrication processes. **c**. The simulated $|E|^2$ profile of the TE_0 mode in a cSb_2Se_3 waveguide, which is 1.2-μm-wide, 30-nm-thick, sits on top of a 330-nm-thick Si_3N_4-on-insulator substrate and is capped with a 200-nm-thick SiO_2 for protection. **d**. Optical image of aSb_2Se_3 resolution test patterns written on cSb_2Se_3 thin film. The minimum feature size achieved is 300 nm. **e**. The same test pattern as in **d** is erased back to cSb_2Se_3.

4.2 Substrate preparation and direct laser writing setup

4.2.1 *Substrate preparation and phase-change PIC parameters*

As shown in Figure 4.2, to prepare the substrate for writing a PIC, only two steps, Sb_2Se_3 sputtering and SiO_2 deposition, are required. The PIC is written on a standard oxidized silicon substrate, which is coated with a 200-nm-thick SiO_2 layer covering a 30-nm-thick Sb_2Se_3 layer on a 330-nm-thick Si_3N_4 film. The SiO_2 capping layer protects and prevents oxidation of the Sb_2Se_3 layer.

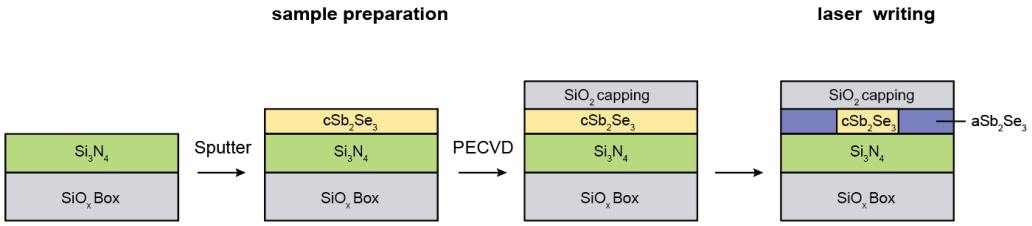

Figure 4.2 Process flows for direct laser writing of PICs. Only two steps, Sb_2Se_3 sputtering and SiO_2 deposition, are required for sample preparation. PECVD: plasma-enhanced chemical vapor deposition.

Table 4.1 provides a concise overview of the key parameters used for phase-change PICs, allowing for clarity.

Table 4.1. Parameters of the phase-change PICs

SiO_2 capping thickness (nm)	200	Sb_2Se_3 thickness (nm)	30
Si_3N_4 thickness (nm)	330	SiO_2 box thickness (μm)	2.8
Sb_2Se_3 grating coupler period (nm)	890	Grating coupler duty cycle	0.6
Sb_2Se_3 waveguide width (μm)	1.2	Waveguide Bending radius (μm)	120
Directional coupler gap distance (nm)	350	Ring resonator radius (μm)	120

4.2.2 *Direct laser writing setup*

We employ a commercial direct writing lithography tool (Heidelberg DWL66+ with its Hires laser head) for the writing of the phase-change PICs (Figure 4.3). We emphasize the advantages of utilizing this commercial laser writing tool as it provides both high writing speed and resolution, enabling efficient and precise fabrication of photonic structures. The tool utilizes a continuous wave (CW) 405 nm diode laser as the laser source. An acousto-optic modulator (AOM) is employed to regulate and modulate the laser intensity. Subsequently, the laser beam passes through an acousto-optic deflector (AOD) and is focused onto the sample surface using a high numerical

aperture (NA) lens. Due to AOD's limited deflection angle, this tool's maximum writing field is 60 μm. To fully extend the patterning range, the sample is placed on a high-speed and high-precision 2D motion stage. By incorporating both the AOD and the motion stage, the system enables the fast writing of PIC design on a phase-change thin film with a minimum feature size of 300 nm. In our experiment, we set the laser power to 27.5 mW, respectively, with a laser scanning speed of 3 mm^2/min. Furthermore, another key benefit of the tool is its compatibility with commonly used patterning software. Specifically, the tool can directly load designs in formats such as GDS or CAD and transfer the design into patterns, eliminating the need for additional pattern conversion procedures.

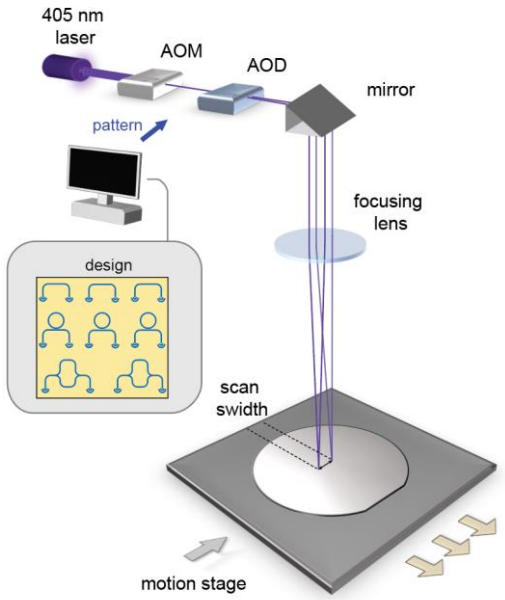

Figure 4.3 Schematic diagram of the direct laser writing experimental setup. AOM: acousto-optic modulator. AOD: acousto-optic deflector.

While using a commercially available direct laser writer is convenient, we note that it is possible to construct a homebuilt laser writing system to minimize equipment requirements. To this end, we also demonstrate the approach that utilizes a cheap, high-power laser diode as the source for laser writing.

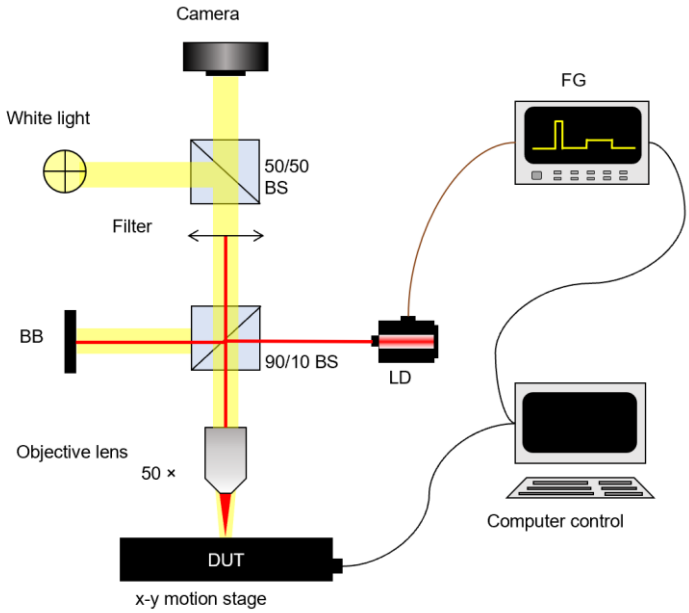

Figure 4.4 Schematic diagram of the homemade direct laser writing setup. LD: laser diode. BS: beam splitter. BB: beam block. FG: function generator. DUT: device under test.

As illustrated in Figure 4.4, a 637 nm laser diode with a maximum output power of 170 mW (HL63133DG) is utilized for the writing and erasing processes. Precise control of laser pulse parameters is achieved by modulating the diode current with a function generator. The modulated laser output is then focused onto the substrate using a 50× objective lens with an NA of 0.55. To write a single spot, a rectangular pulse of 200 ns pulse duration and pulse energy of 50 nJ is used. To erase a single spot, the pulse duration is 5 ms with a pulse energy of 333 µJ. Both the writing and erasing processes achieved a diffraction-limited resolution of 500 nm, which can be further improved using a laser diode with a shorter wavelength. To facilitate precise positioning of the substrate, a 2D x-y closed-loop motion stage is used, which is capable of achieving a minimum moving step size of 50 nm. By combining the controlled light pulse with the movement of the substrate on the stage, the desired pattern was achieved. Moreover, for faster erasure of the pattern over a larger area, the laser diode was operated in continuous wave (CW) mode at a fixed power of 15 mW while scanning the stage at a speed of 0.1 mm^2/min. This method effectively erased all patterns in the path of the laser spot.

We note that the writing of PICs is performed using the commercial Heidelberg DLW66+ writer, while most of the local tuning and erasing of the PICs is achieved by the homebuilt system.

4.3 Phase-change integrated photonic elements simulation and characterization

We first demonstrate the DLW of high-quality building-block components of PICs. The characterization of photonic building block elements, including directional couplers, waveguide crossings, inverse-designed bends with low loss, and so on, are included in this section.

4.3.1 *Directional couplers*

We write directional couplers with various coupling lengths on a phase-change thin film for characterization purposes. The cSb_2Se_3 waveguide used in the couplers has a width of 1.2 μm, and the gap between the coupled waveguides is 350 nm. The measured split ratio between the bar port and the cross port (defined as shown in Figure 4.5a) aligns well with the simulation results. The split ratio of the directional coupler can be tuned by increasing the gap between the coupled waveguides from 350 nm to 500 nm over a distance of tuning length (Figure 4.5b). Fig. S5d shows the simulation result of the split ratio between the bar and cross ports when changing the tuning length of the directional coupler with a total coupling length of 90 μm. Importantly, this reduction of the waveguide width does not introduce any significant insertion loss (<0.5 dB).

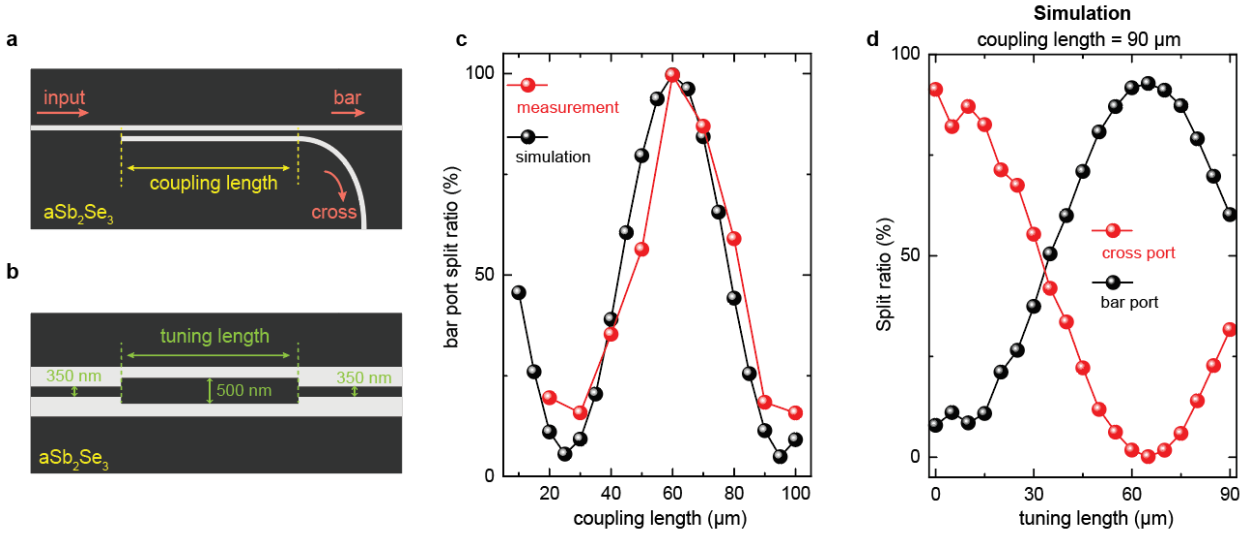

Figure 4.5 Simulation and experimental characterization of the Sb_2Se_3 directional coupler. **a.** Schematic illustration of the directional coupler calibration measurement. **b.** The zoomed-in schematic of the coupling region showing the coupling of the directional coupler can be tuned by increasing the gap between waveguides from 350 nm to 500 nm. **c.** The bar port transmission ratio

is a function of the coupling length. **d**. The split ratio of the directional coupler when tuning the erase length.

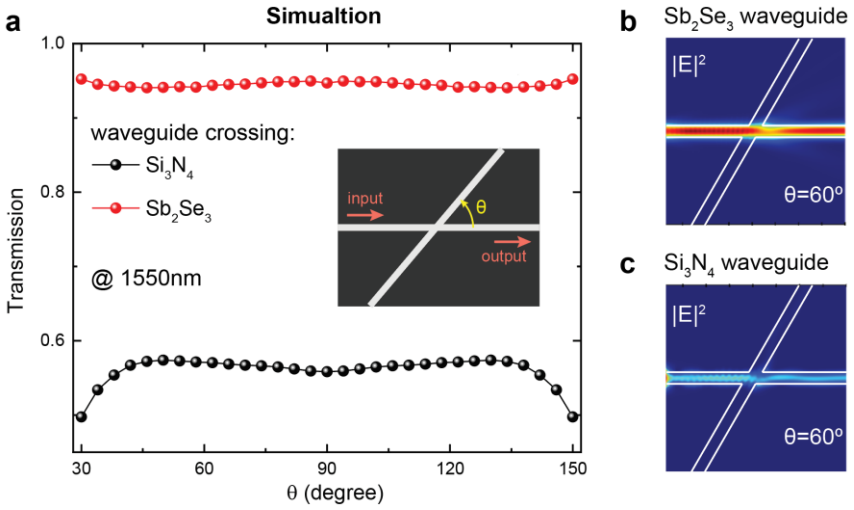

Figure 4.6 **a.** Simulation of the transmission of the Sb_2Se_3 and Si_3N_4 waveguide crossings. Inset: the schematic of the waveguide crossing geometry. **b** and **c.** The $|E|^2$ distribution of the Sb_2Se_3 waveguide crossing (**b**) and Si_3N_4 waveguide crossing (**c**) when $\theta = 60°$. The transmission of the Sb_2Se_3 waveguide crossing is much higher than the Si_3N_4 waveguide crossing.

4.3.2 *Waveguide Crossing*

In conventional dielectric waveguides, such as Si_3N_4 waveguides (e.g. 1.2 μm wide, 330 nm thick) on a SiO_2 substrate, crosstalk between waveguides becomes non-negligible when they cross over each other. As a result, special waveguide crossing designs are required to minimize the crosstalk and maintain signal integrity. In the case of phase-change waveguides (e.g. 1.2 μm wide, 30 nm thick cSb_2Se_3 on top of a Si_3N_4 substrate), the crosstalk is significantly reduced even when the two waveguides directly cross over each other over a wide range of cross angles. This is because the phase-change waveguide has a weaker confinement compared to the conventional dielectric waveguide. This makes the optical mode in the Sb_2Se_3 waveguide much broader and experiences a weaker index contrast at the waveguide crossing region. Consequently, no additional design modifications are necessary to mitigate the crosstalk in these waveguide crossings, as shown in our simulation results in Figure 4.6.

4.3.3 *Inverse-designed waveguide bends*

The cSb_2Se_3 waveguide is more sensitive to bending losses compared to conventional dielectric waveguides. This is attributed to the weaker mode confinement and additional scattering caused by randomly oriented crystalline grains. To mitigate bending losses in our photonic devices, we set the bending radius to 120 μm. To further reduce bending losses and enable tighter bending radius, we utilized an inverse-design method to design waveguide bend structures. As depicted in Figure 4.7a, these inverse-designed waveguide bends incorporate a DBR-like structure outside the bending region, which helps decrease the loss.

We write multiple pairs of single-mode phase-change waveguides, each pair consisting of one waveguide with inverse-designed waveguide bends and another with conventional arc-shape waveguide bends. Both designs featured a bending radius of 35 μm. Our measurements confirm a significant reduction in bending loss by two orders of magnitude when using inverse-designed waveguide bends. It is also worth noting that the conventional arc-shape waveguide bends with a radius of 35 μm exhibited higher bending losses (>10 dB per bend) in the measurement compared to the simulation results (< 0.5 dB per bend). This discrepancy may be attributed to scattering effects caused by the presence of grains within the waveguide, which is not fully captured in the simulations.

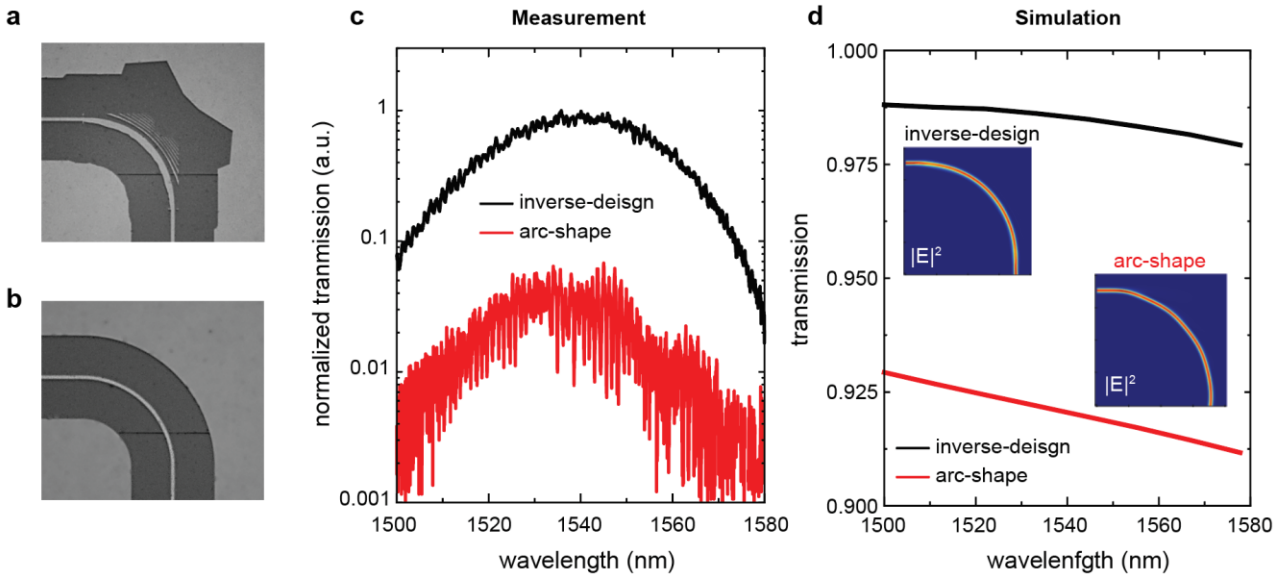

Figure 4.7 **a** and **b.** Optical images of the inverse-designed (**a**) and arc-shape (**b**) waveguide bends written on phase-change thin film, respectively. Both waveguide bends have a bending radius of 35 μm. **c.** The measured transmission spectra of single mode waveguide consist of one waveguide and two waveguide bends. **d.** The simulated transmission spectra of the waveguide bend. Inset: the mode profile of the inverse-designed waveguide bend and the arc-shape bend at 1550 nm.

4.3.4 *Racetrack ring resonator and Mach-Zender interferometer*

We first demonstrate the DLW of high-quality building-block components of PICs. The first example (Figure 4.8a) is a racetrack ring resonator coupled with a bus waveguide connected to input and output grating couplers, all directly written with DLW. The resonator has a width of 1.2 μm and a radius of 120 μm with a gap of 350 nm to the bus waveguide (Figure 4.8b). The pair of grating couplers couple light from optical fibers into the bus waveguide (Figure 4.8c) with an efficiency of -14 dB. Figure 4.8d shows the measured transmission spectrum through the bus waveguide, showing the resonances of the ring resonator with an intrinsic quality factor is $Q_i \sim$ 12700, corresponding to a propagation loss of 2.8 dB·mm^{-1}.

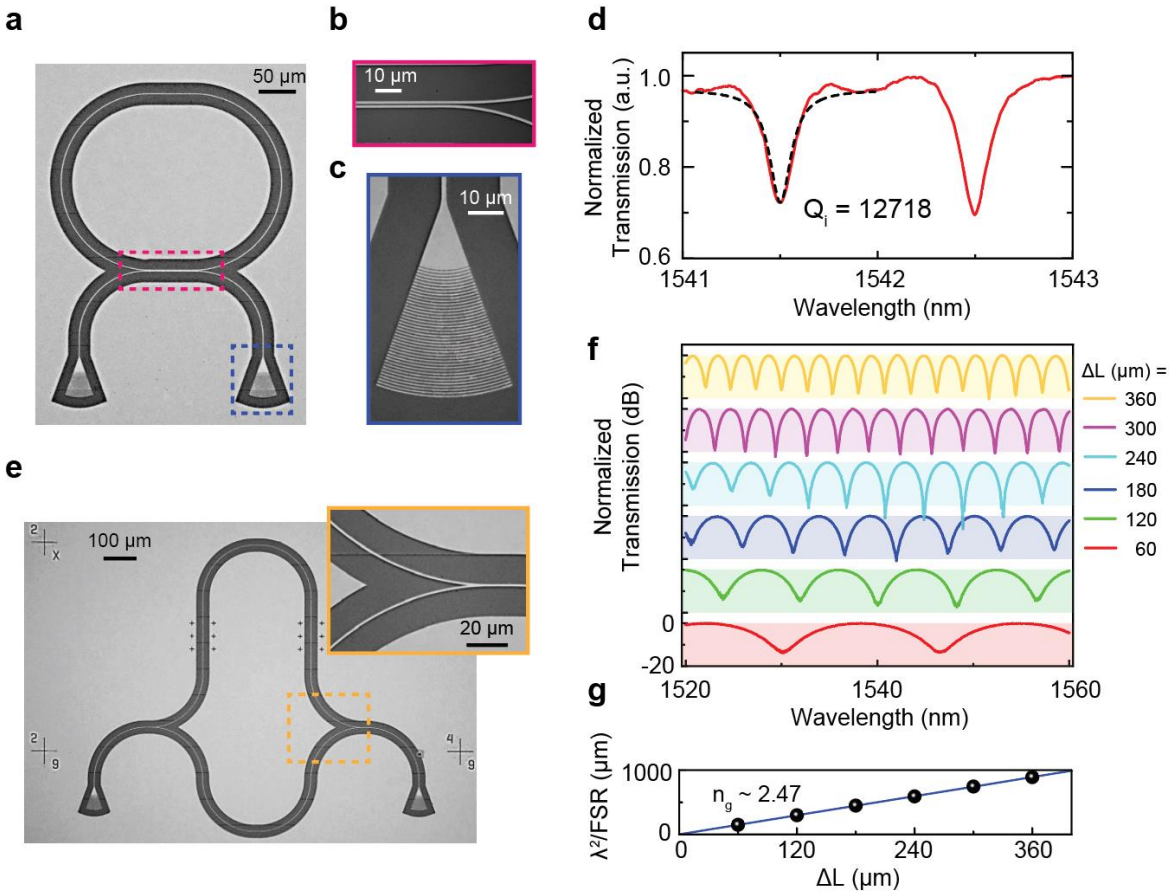

Figure 4.8 Direct-write photonic components and their characterization. **a**. Optical image of a racetrack ring resonator that is directly written on the Sb_2Se_3 thin film. **b** and **c**. The zoomed-in optical image of the waveguide-ring coupling region (**b**) and the grating coupler (**c**). **d**. The transmission spectrum of the ring resonator is shown in (**a**). The spectrum is normalized to the spectrum of a pristine Sb_2Se_3 waveguide. The intrinsic Q factor is 12718. **e**. Optical image of a direct-write PCM Mach-Zehnder interferometer (MZI). Inset: the zoomed-in image of the Y-combiner part in the MZI. **f**. Normalized transmission spectra of the MZIs. Spectra have been vertically offset for clarity**. g**. λ^2/free-spectral-range vs. the length difference between two arms of the MZI, ΔL. Linear fitting yields the waveguide group index n_g to be 2.47.

The loss in the cSb_2Se_3 waveguide is mainly attributed to the scattering induced by the grain boundaries in the waveguide, as in-plane crystalline grains within the cSb_2Se_3 region can be observed with high-resolution electron microscopy. These grains, typically a few microns in size, are also randomly oriented. Because of the optical birefringence of cSb_2Se_3, these randomly oriented grains result in the uneven absorption of laser energy during the DLW and limit the smoothness of the waveguide. The scattering loss can be mitigated when a lower erasure temperature is used to form larger crystal grains with sizes on the order of tens of microns, which

we have observed. Further mitigation of the loss will require improved optical phase-change materials with finer polycrystalline structures. Nevertheless, the propagation loss of the cSb_2Se_3 waveguide is consistent with the reported values of dielectric waveguides integrated with Sb_2Se_3.

The devices written on phase-change thin film exhibit consistent and reliable performance. To demonstrate this, we write multiple Mach-Zehnder interferometers (MZIs) with varied path length differences, ΔL, between the two arms (Figure 4.8e). The measured transmission spectrum is shown in Figure 4.8f. The consistently high extinction ratio (> 20 dB) indicates the 50/50 beam splitters (inset, Figure 4.8e) perform very well. Fitting the spectra, we extract a group index n_g = 2.47 for a 1.2-μm-wide Sb_2Se_3 waveguide, which agrees well with the simulated value of 2.43.

4.3.5 *Local tuning of the phase-change photonic waveguide*

The transmission of the Sb_2Se_3 waveguide can be tuned by selectively erasing or rewriting specific portions of the waveguide. Erasing a section of the waveguide results in a decrease in transmission at the output. Conversely, rewriting a section of the waveguide leads to an increase in transmission (Figure 4.9a and Figure 4.9b). To demonstrate this transmission tuning capability, we erased a 1.2 μm wide waveguide in incremental steps to create a gap up to 30 μm. The corresponding transmission spectra displayed a consistent global decrease in transmission, as illustrated in Figure 4.9b and Figure 4.9c. Subsequently, we restored the erased portion of the waveguide, leading to the recovery of the original transmission spectrum and its associated transmission level.

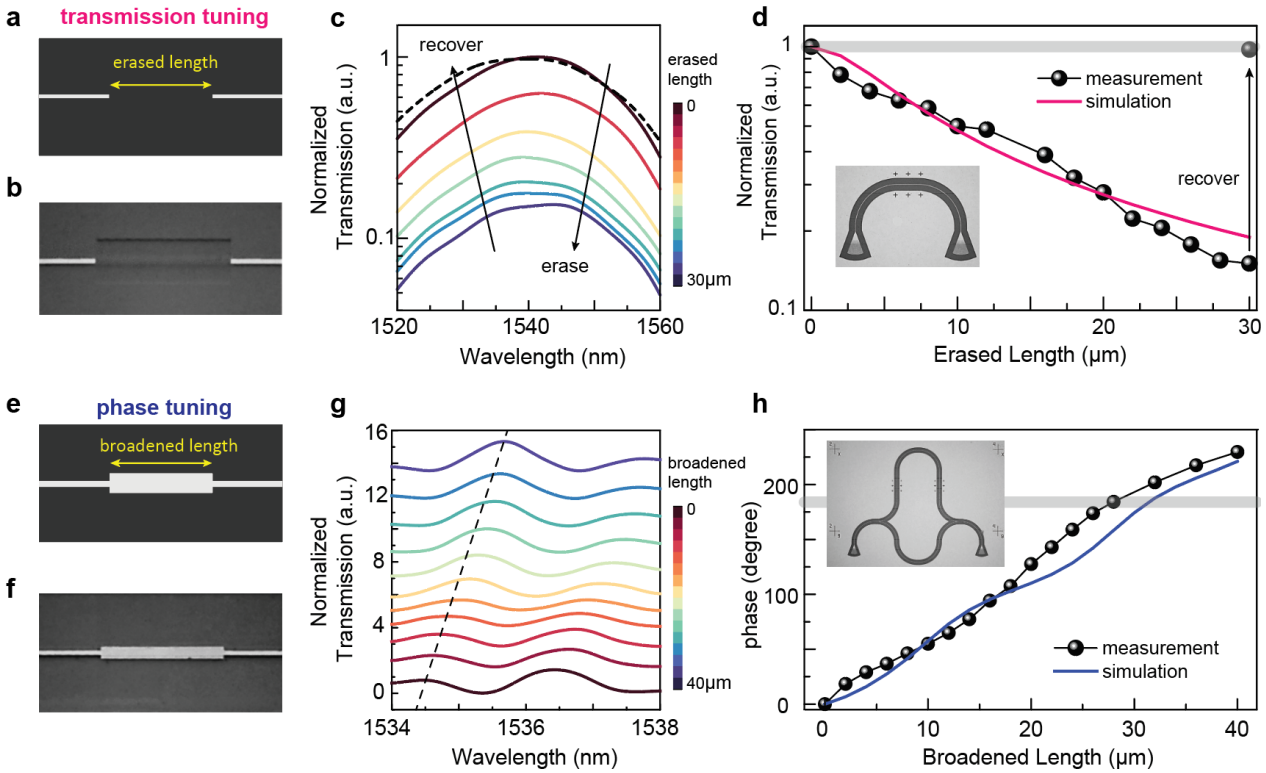

Figure 4.9 Tuning the transmission and the phase response of a Sb_2Se_3 waveguide. **a.** Schematic of tuning the transmission by varying the erased waveguide length. **b.** Optical image of a partially erased Sb_2Se_3 waveguide. **c.** The transmission spectrum of a Sb_2Se_3 waveguide when an erased length increased from 0 μm to 30 μm in steps. **d.** The waveguide transmission (normalized to the transmission of the pristine Sb_2Se_3 waveguide) under different erased lengths. Inset: the optical image of the waveguide used for transmission measurement. **e.** Schematic of tuning the phase response of a Sb_2Se_3 waveguide by broadening waveguide width. **f.** Optical image of a Sb_2Se_3 waveguide where the center part is broadened. **g.** The shift of transmission spectrum of an MZI when quasi-continuously changing the length of the broadened region in one arm of the MZI from 0 μm to 40 μm. **h.** The induced phase shift under different broadened waveguide lengths. Inset: the optical image of the MZI used for phase-tuning measurement.

In addition to transmission tuning, the phase response of the waveguide can be selectively tuned by widening specific sections of the waveguide from 1.2 μm to 2 μm (Figure 4.9e and Figure 4.9f). We performed transmission spectra measurements after introducing phase shifts in one arm of the MZI. As increasing the length of the widened waveguide in incremental steps, the interference pattern experienced a gradual redshift, as depicted in Figure 4.9g. Through fitting

analysis of the measurements (Figure 4.9f), we determined that a length of 30 μm of the widened waveguide from 1.2 μm to 2 μm can achieve a π phase shift.

4.4 Examples of direct-write & rewritable PIC components

In this section, we demonstrate the versatility of this DLW technique with various photonic circuits for diverse applications, including an optical interconnect fabric for reconfigurable networking, a photonic crossbar array as a tensor core for optical computing, and a tunable optical filter for optical signal processing.

4.4.1 *Programmable interconnect fabric and crossbar matrix*

Programmable optical interconnect fabric is a crucial photonic architecture to enable reconfigurable connectivity in telecommunication, computing, and network systems. Using the DLW technique, we can create and, on-demand, route and reroute such interconnect fabric. In Figure 4.10a to c, we demonstrate a 3×3 array, which establishes connectivity between three input and output ports using cSb_2Se_3 waveguides. We define the connection configuration of the fabric using a transmission matrix **M** in $S_{out}=\mathbf{M}\cdot S_{in}$, where S_{out} and S_{in} represent the output and input intensity vectors, respectively. As shown in Figure 4.10a, we initially patterned the fabric to connect input/output ports in a configuration of 1→2, 2→1, and 3→3. The measurement result of this transmission matrix **M** shows that the extinction ratio between desired and undesired connections exceeds 20 dB (Figure 4.10d). We then reroute the fabric by erasing all the waveguides in the designated connection region (Figure 4.10b) and subsequently rewriting the waveguides to establish new connections with an updated configuration. As illustrated in Figure 4.10c, we reroute the switch array to the configuration of 1→2, 2→3, and 3→1. The measured **M** has changed accordingly but maintains a high extinction ratio >18 dB. Additionally, we observe a residual image from the previous writing even after completely erasing the pattern. The residual image does not impact the performance of the newly patterned devices, though. We argue that this

residual imprint can be explained by the optical birefringence resulting from size and orientation variations in cSb$_2$Se$_3$ crystal grains during the annealing processes. Specifically, during the laser erasing process, the crystal grains grow in alignment with the moving direction of the laser spot, while no preferred growth direction is present during the thermal erasing process.

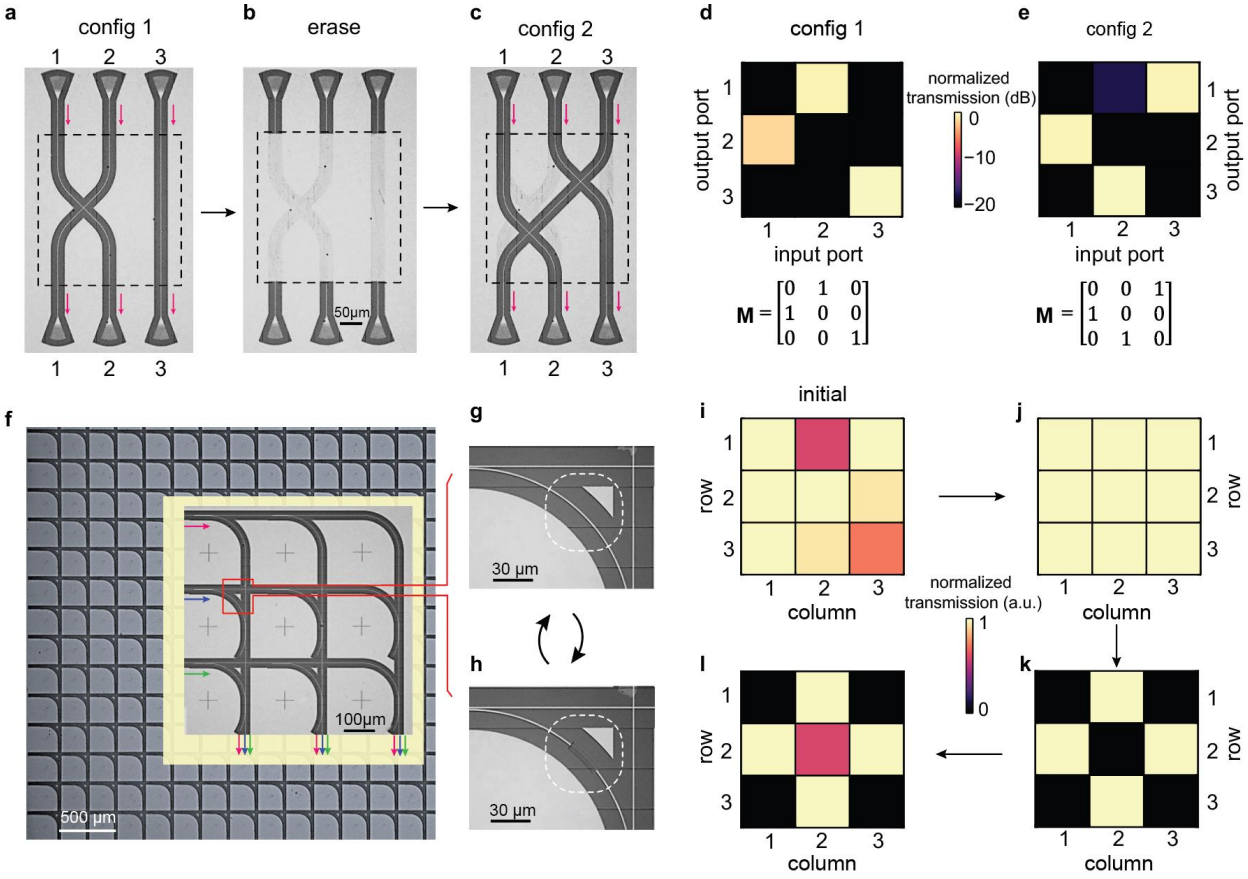

Figure 4.10 Programmable photonic switch array and crossbar array. **a.** Optical image of a 3×3 optical switch array with the initial connection configuration (matrix 1). The connection region of the optical switch is then erased (**b**) and rewritten (**c**) into a new connection configuration (matrix 2). **d** and **e.** The measured transmission matrix of switch configuration 1 (**d**) and configuration 2 (**e**), respectively. **f.** Optical image of a DLW 14×14 crossbar array. Inset: the 3×3 sub-array used for testing. To decrease or increase the transmission of a specific crossbar element, the cross-coupling waveguide is partially erased (**h**) or recovered (**g**), respectively. **i-k.** The transmission matrix of the crossbar array after each configuration step. The crossbar array is initially designed to equally distribute the input power into outputs. Due to writing imperfections, the transmission matrix of the crossbar array has errors (**i**) and is corrected by DLW to restore the designed functionality (**j**). **k and l.** Transmission matrix after setting the element at the center and four corners to 0 (**k**) and then resetting the center matrix element to 0.5 (**l**).

A more complicated transmission matrix can be achieved by using a photonic crossbar array, which is a universal architecture for implementing MAC operations in linear optical computing. Figure 4.10f displays the optical image of a 14×14 photonic crossbar array written on the PCM thin film. As proof of concept, we demonstrate the operation of a 3×3 sub-array to represent a programmable MAC core. The input vector is encoded by the intensities of optical signals in multiple wavelength channels, which are sent into each row of the crossbar array, respectively. As illustrated in Figure 4.10g, each unit cell of the crossbar array consists of a row waveguide, a column waveguide, and a cross-coupling waveguide, which extracts the input signal from the row waveguide and couples it to the column waveguide through optimized directional couplers. The output of each column waveguide is the weighted sum of the input signal at each row and is measured incoherently by a photodetector, thus performing the MAC operation for MVM. Each matrix element is programmed in the transmission of each unit cell's cross-coupling waveguide (see Figure 4.10f). The output vector is obtained by measuring the total output power in each column of the output channels. A more detailed elucidation of crossbar array implementation can be found in Chapter 2.5.

For demonstration, we initially designed and direct-wrote the 3×3 crossbar array to equally distribute the input power into the output to represent the matrix $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$. However, due to writing imperfections (similar to fabrication variation in conventional PICs), the matrix represented by the crossbar array exhibits errors, as shown in the measured results in Figure 4.10i. Remarkably, the DLW technique makes it straightforward to adjust each element by locally modifying the cross-coupling waveguides, thereby correcting the error to restore the designed functionality (Figure 4.10j). It is also easy to reprogram the matrix in a grayscale-like fashion by erasing or restoring a section of the cross-coupling waveguide. As an example, we can set the element at the center and four corners to 0 (Figure 4.10k) and decrease the center matrix element to 0.5 (Figure 4.10l) to represent the matrix $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0.5 & 1 \\ 0 & 1 & 0 \end{bmatrix}$.

4.4.2 *Shaping Spectral Response of Coherent PICs*

Coherent PICs universally use interferometric and resonant components and require precise spectral responses for filtering and coherent signal processing [132,133]. Because of inevitable fabrication imprecision, it is necessary to shape these components' spectral responses to meet these requirements. The DLW technique allows us to incrementally trim the spectral response of a coherent optical filter step-by-step, offering a highly controlled method for fine-tuning its performance. In addition, we can even add new coherent components to change the spectral response entirely. As shown in Figure 4.11a, we first write an MZI, which has a typical transmission spectral response as measured in Figure 4.11g. We then erase a section of the lower arm of the MZI (Figure 4.11b) and subsequently restore it and add a racetrack ring resonator, as depicted in Figure 4.11c. The resulting transmission spectrum of the ring coupled-MZI displays features the characteristics of both the MZI and the ring resonator. Afterward, we erase portions of both the ring resonator and the MZI (see Figure 4.11d) and rewrite the circuit to a double-injection ring (DIR) filter [134]. In the DIR, the output light consists of cumulative contributions from two nominally identical add-drop ring resonators. The first contribution comes from the drop-port of the ring, with the input light incident from the lower arm of the original MZI. The second contribution originates from the through-port of the ring, with the light incident from the upper arm of the original MZI. As a result, the DIR exhibits a significantly different transmission spectrum with a much larger FSR (Figure 4.11g). Finally, we can trim the DIR spectral response by modifying its parameters. We adjust the coupling between the output waveguide and the racetrack ring resonators. This is done by widening the gap between the waveguides from 350 nm to 500 nm and reducing the waveguide width (Figure 4.11f). The transmission spectrum of the modified DIR (DIR2 in Figure 4.11f) changes as expected, thus demonstrating the spectral tunability of the filter. We note that all measured spectra from the filter agree well with the numerical models. This provides additional evidence of the stability and reliability of writing and rewriting the phase-change photonic circuits.

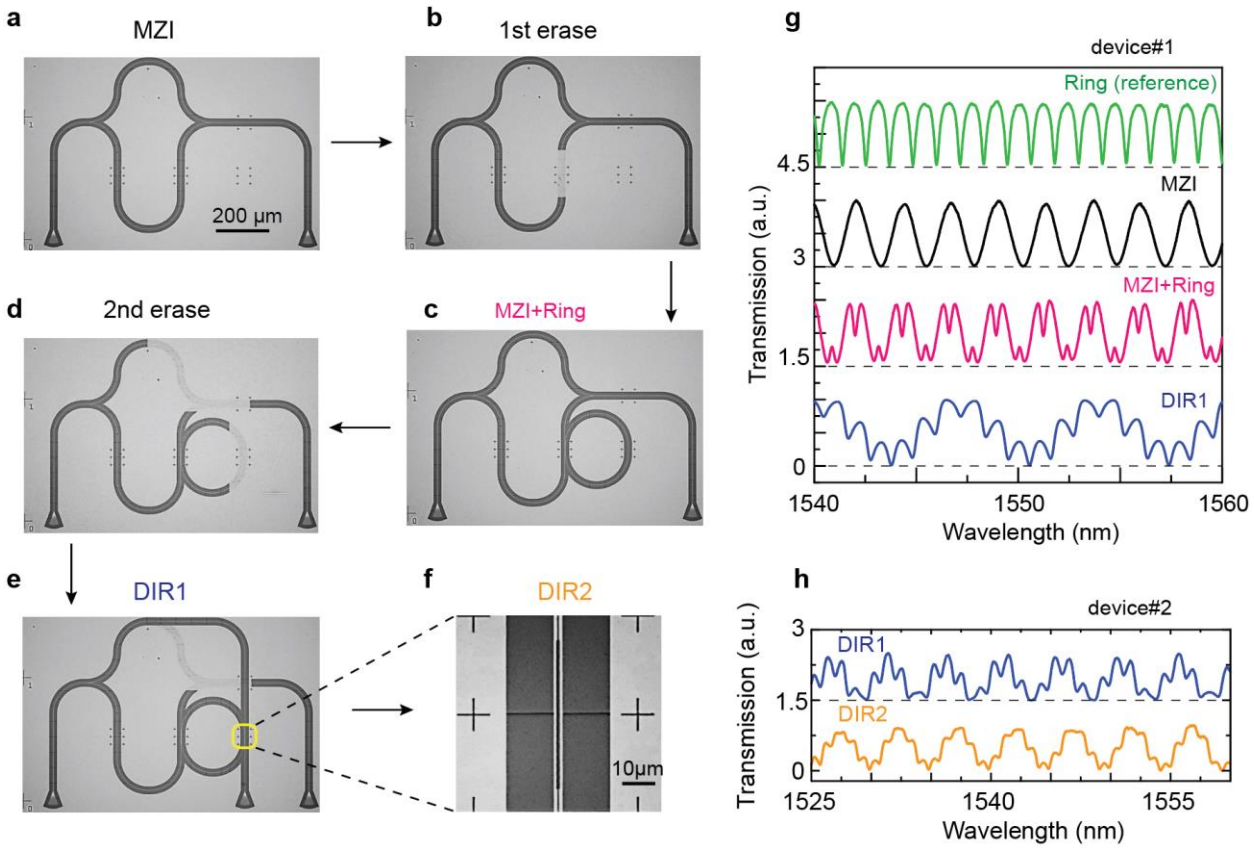

Figure 4.11 Shape the spectral response of an optical filter in steps. **a.** An MZI was written by DLW on a Sb_2Se_3 thin film. **b.** Erase part of the bottom arm of the MZI. **c.** Add a ring resonator to the erased region of the MZI. **d.** Erase part of the ring resonator and the MZI. **e.** Reconnect the device as the double-injection ring (DIR) filter. **f.** Tuning the coupling by increasing the gap between the ring resonator and the top arm of the Y-splitter. **g.** The transmission spectra of the optical filter after each reconfiguration step, respectively. Spectra have been vertically translated for clarity. **h.** The transmission spectra of the double-injection ring filter before (blue curve) and after (orange curve) tuning the ring resonator coupling.

4.4.3 *Modeling the tunable optical filter*

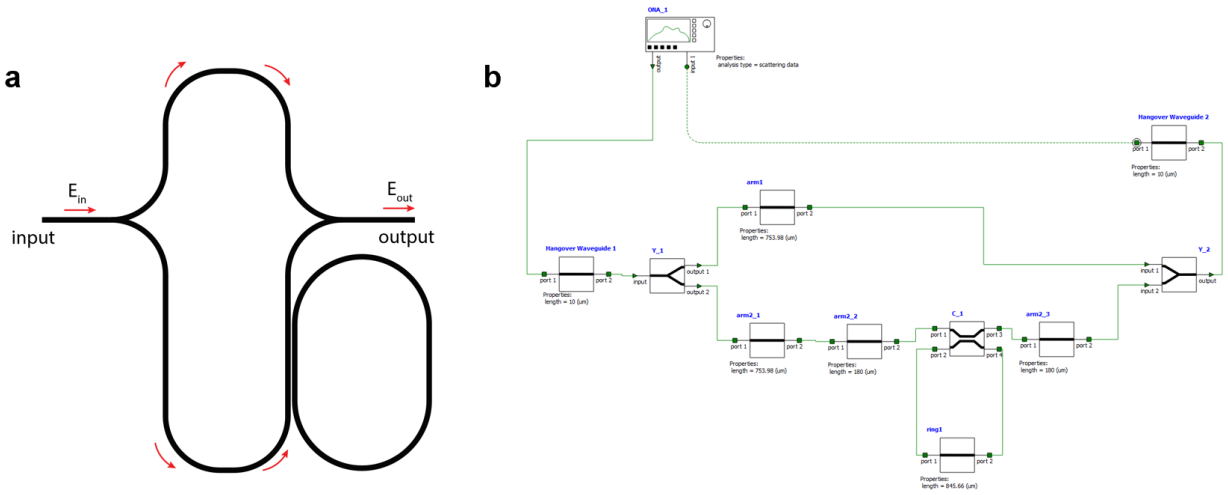

Figure 4.12 The schematic of the ring resonator coupled MZI (**a**) and the ANSYS INTERCONNECT model which is used to simulate the device's performance (**b**).

The performance of the optical filter composed of a ring resonator added to an MZI is simulated using ANSYS INTERCONNECT software (Figure 4.12). Two devices with different parameters are written and measured experimentally. The response of the devices with the same geometry is also simulated. The dimensions of the devices used in the simulation are the same as the design. Other parameters used in the simulation including the propagation loss, the group index, and so on are obtained from the experimental device characterization. The parameters are listed in Table 4.2. As shown in Figure 4.13, the simulation results align with the experimental measurement.

Table 4.2. Parameters of ring-coupled MZIs

Device #1		Device #2	
Sb_2Se_3 waveguide width (μm)	1.2	Sb_2Se_3 waveguide width (μm)	1.2
Effective index n_{eff}	1.91	Effective index n_{eff}	1.91
Group index n_{group}	2.47	Group index n_{group}	2.47
Propagation loss (dB/mm)	4.5	Propagation loss (dB/mm)	4.5
MZI arm difference ΔL (μm)	420	MZI arm difference ΔL (μm)	420
Ring resonators length L_{Ring}(μm)	845.6	Ring resonators length L_{Ring}(μm)	955.7
Ring & waveguide coupling coefficient	0.45	Ring & waveguide coupling coefficient	0.9

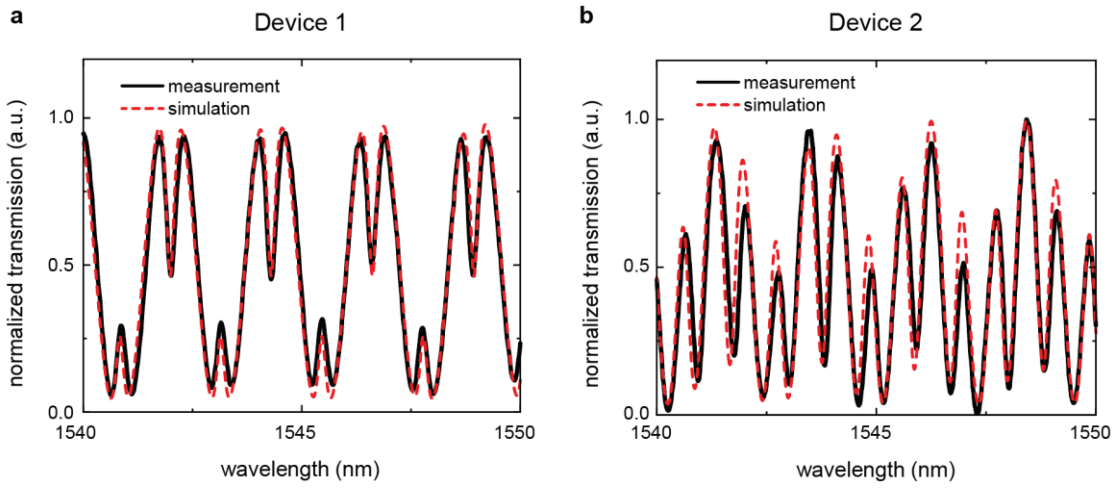

Figure 4.13 The simulation results of two different filters (a ring resonator added to the MZI) with different parameters.

Both two filters are then rewritten to perform as the DIR with the geometry defined in Figure 4.14 Schematic illustration of a double injection ring filter. Mathematically, the model describing the transmitted electromagnetic field dependence on the wavelength of the DIR is:

$$E_{t1}(\lambda)=\frac{\left(\tau_1-\tau_2^{*}\alpha e^{-i\theta}\right)}{1-\tau_1^{*}\tau_2^{*}\alpha e^{-i\theta}}\left|E_{i1}(\lambda)\right|e^{-i\phi_{i1}}-\frac{\kappa_1\kappa_2^{*}\sqrt{\alpha}e^{-i\frac{\theta}{2}}}{1-\tau_1^{*}\tau_2^{*}\alpha e^{-i\theta}}\left|E_{i2}(\lambda)\right|e^{-i\phi_{i1}} \tag{4.1}$$

where $\tau=|\tau|e^{-i\varphi_\tau}$, $\kappa=|\kappa|e^{-i\varphi_\kappa}$, α, E_i, and Φ_i are the transmission and coupling coefficients of the directional couplers, the loss coefficient of the ring, and the injected fields' intensity and their phase, respectively. θ is the accumulated phase as the light traverses the rg at a steady state:

$$\theta(\lambda)=\frac{2\pi}{\lambda}n_{eff}(\lambda)L_{Ring} \tag{4.2}$$

where n_{eff} is the effective index of the propagating mode and L_{Ring} is the perimeter of the ring resonator.

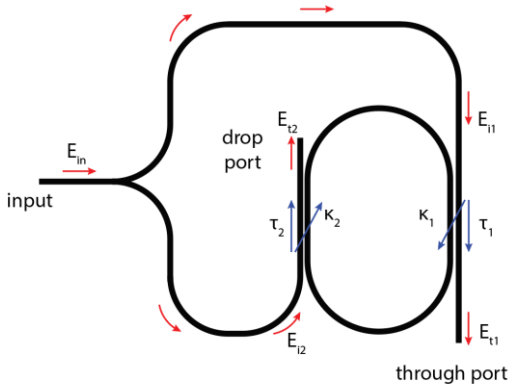

Figure 4.14 Schematic illustration of a double injection ring filter.

The spectral responses of the DIR devices are measured and well-fitted, as shown in Figure 4.15 and Figure 4.16. After measurement, we further tune the spectral responses of the DIR devices by modifying the transmission and coupling coefficients of the two DIRs, respectively. Specifically, we tuned the transmission coefficient τ_2 and coupling coefficient κ_2 in the first DIR (device #1) and tuned the transmission coefficient τ_1 and coupling coefficient κ_1 in the second DIR (device

#2). The spectra responses of the DIRs change accordingly (see Figure 4.15b and Figure 4.16b). The parameters are listed in Table 4.3 and Table 4.4.

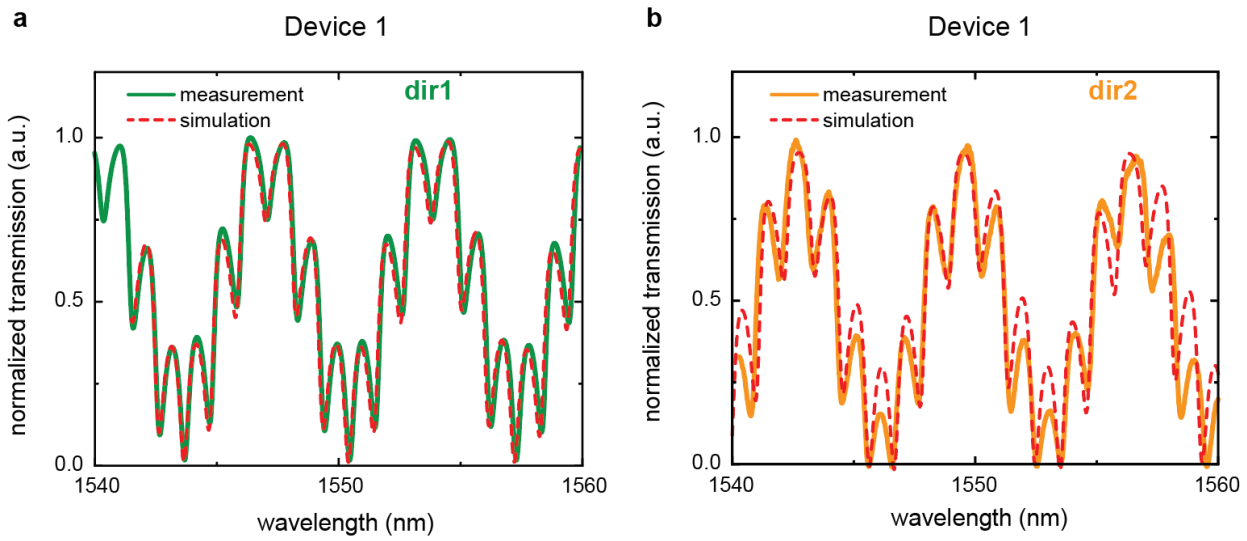

Figure 4.15 **a.** Measured and simulated spectral response of the first double injection ring filter, device #1. **b.** After tuning the transmission coefficient τ_2 and coupling coefficient κ_2, the spectral response of the same device (DIR2) changes.

Table 4.3. Fitting Parameters of DIR Device #1

Before tuning (DIR1)		After tuning (DIR2)	
Coupling coefficient 1 $\lvert\kappa_1\rvert^2$	0.45	Coupling coefficient 1 $\lvert\kappa_1\rvert^2$	0.45
Coupling phase 1 $\Phi_{\kappa 1}$ (degree)	0	Coupling phase 1 $\Phi_{\kappa 1}$ (degree)	0
Coupling coefficient 2 $\lvert\kappa_2\rvert^2$	0.45	Coupling coefficient 2 $\lvert\kappa_2\rvert^2$	0.4
Coupling phase 1 $\Phi_{\kappa 2}$ (degree)	0	Coupling phase 1 $\Phi_{\kappa 2}$ (degree)	30
Loss coefficient of ring α	0.45	Loss coefficient of ring α	0.45
Injected field intensity 1 $\lvert E_{i1}\rvert^2$	50%	Injected field intensity 1 $\lvert E_{i1}\rvert^2$	50%
Injected field intensity 2 $\lvert E_{i2}\rvert^2$	50%	Injected field intensity 2 $\lvert E_{i2}\rvert^2$	50%

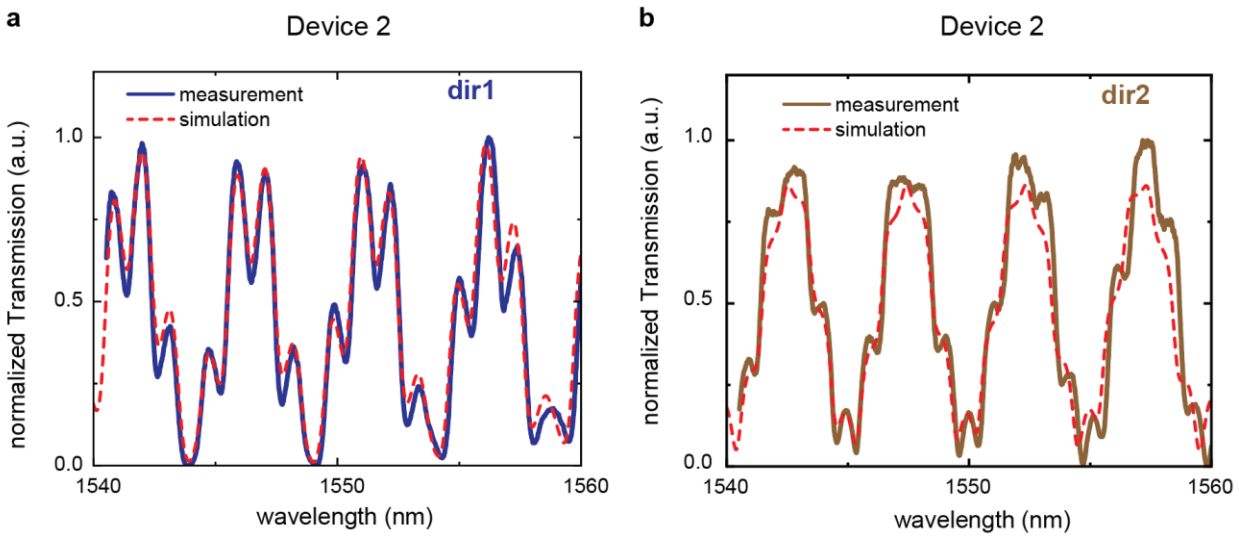

Figure 4.16 **a.** Measured and simulated spectral response of the second double injection ring filter, device #2. **b.** After tuning the transmission coefficient τ_1 and coupling coefficient κ_1, the spectral response of the same device (DIR2) changes.

Table 4.4. Fitting Parameters of DIR Device #2

Before tuning (DIR1)		After tuning (DIR2)	
Coupling coefficient 1 $\lvert\kappa_1\rvert^2$	0.9	Coupling coefficient 1 $\lvert\kappa_1\rvert^2$	0.22
Coupling phase 1 $\Phi_{\kappa 1}$ (degree)	0	Coupling phase 1 $\Phi_{\kappa 1}$ (degree)	30
Coupling coefficient 2 $\lvert\kappa_2\rvert^2$	0.9	Coupling coefficient 2 $\lvert\kappa_2\rvert^2$	0.9
Coupling phase 1 $\Phi_{\kappa 2}$ (degree)	0	Coupling phase 1 $\Phi_{\kappa 2}$ (degree)	0
Loss coefficient of ring α	0.45	Loss coefficient of ring α	0.45
Injected field intensity 1 $\lvert E_{i1}\rvert^2$	50%	Injected field intensity 1 $\lvert E_{i1}\rvert^2$	50%
Injected field intensity 2 $\lvert E_{i2}\rvert^2$	50%	Injected field intensity 2 $\lvert E_{i2}\rvert^2$	50%

4.5 Conclusion and outlook

In conclusion, in this chapter, we have presented a flexible, reliable, and cost-effective technique for directly writing and rewriting photonic circuits on low-loss phase-change thin films. This technique simplifies the complex nanofabrication processes typically required for fabricating integrated photonic devices down to one-step laser writing and enables reusing the same chip/die. Although we have utilized a commercial laser writer in this work, we emphasize that the same results can be achieved with a much lower-cost laser writing system, which incorporates a laser diode, a focusing objective, a high-precision motion stage, and a computer control system. This approach leverages the versatility of the DLW and the nonvolatile, low-loss, and high-contrast properties of the phase-change material, offering an unprecedented level of flexibility. We have demonstrated PICs consisting of a full package of elementary photonic components, including waveguides, grating couplers, ring resonators, MZIs, programmable optical switch fabrics, reconfigurable photonic crossbar array, and tunable optical filters. Although our demonstrations have been conducted in a near-*in-situ* fashion—in steps of writing, measuring, modifying, and checking—real-time reconfiguration and feedback-controlled adaptation of the PICs are entirely feasible [84,135].

Furthermore, the application scenario can be expanded by introducing a multi-level grayscale design [35,63] instead of the current binary design or by selecting appropriate substrates tailored to specific applications. For example, the phase-change thin films can also be integrated on $LiNbO_3$-on-insulator substrates, enabling the development of programmable electro-optical or acoustic-optical circuits. These advantages underscore the potential of the DLW technique in enabling the rapid prototyping and testing of innovative photonic circuits, using only a low-cost tool that is affordable to a wide range of research and education communities.

Chapter 5. Summary and outlook

5.1 Summary of the thesis

Programmable photonic devices that utilize phase change materials offer a range of advantages and applications due to their unique properties. PCMs enable the reconfiguration of photonic devices, allowing for dynamic changes in their optical properties. This reconfigurability is particularly valuable in various applications, such as optical switches, modulators, and routers, where the ability to change the device's behavior without physically altering its structure is essential. These devices have the potential for low power consumption. PCMs can switch between states with relatively low energy inputs, making them energy-efficient compared to some other conventional materials used in photonic applications. Meanwhile, programmable photonic devices using phase change materials can be integrated into small footprints, enabling the development of compact and miniaturized photonic circuits. This integration is beneficial for applications where space is limited, such as in integrated photonics for on-chip communication and sensing. The unique properties of programmable photonic devices using phase change materials open up possibilities for novel applications in areas such as reconfigurable optical networks, all-optical signal processing, photonic memories, and neuromorphic computing.

However, challenges like stability, endurance, and scalability still need to be addressed for broader commercial adoption of these devices. Despite these challenges, the development of programmable photonic devices using phase change materials holds great promise for advancing optical technologies with enhanced functionality, speed, and efficiency. In Chapter 2, we show the creation of a versatile photonic computing core employing programmable mode converters using phase-change materials, achieving precise control over spatial modes for image processing and recognition tasks with high accuracy. This core shows potential for large-scale photonic neural networks, boasting ultra-high computation throughputs. In Chapter 3, the focus shifts to mitigating and leveraging noise in photonic computing systems. The development includes a photonic generative network, successfully generating handwritten numbers and demonstrating resilience to hardware imperfections through noise-aware training strategies. In Chapter 4, we introduce direct-write and rewritable photonic circuits based on low-loss phase-change materials. These circuits enable the creation of end-to-end functional photonic circuits in a single step using laser writing,

allowing for easy modification and reprogramming. Various applications, such as optical computing and tunable optical filters for signal processing, showcase the flexibility and versatility of this direct-write technique.

5.2 Future perspectives

The surge in demand for artificial intelligence, high-performance computing, aerospace data collection, hyper-scale data centers, and disaggregated memory and storage has spurred the creation of robust system-on-chip (SoC) solutions, redefining conventional computing architectures. These cutting-edge architectures are designed to amplify data throughput, enhancing communications across dies, sockets, boards, systems, and racks. Over the past quarter-century of Moore's Law scaling, electrical I/O has been a critical bottleneck hindering the scalability of processor and SoC performance. Presently, it stands as a significant barrier impeding the scaling of high-throughput SoC architectures [69].

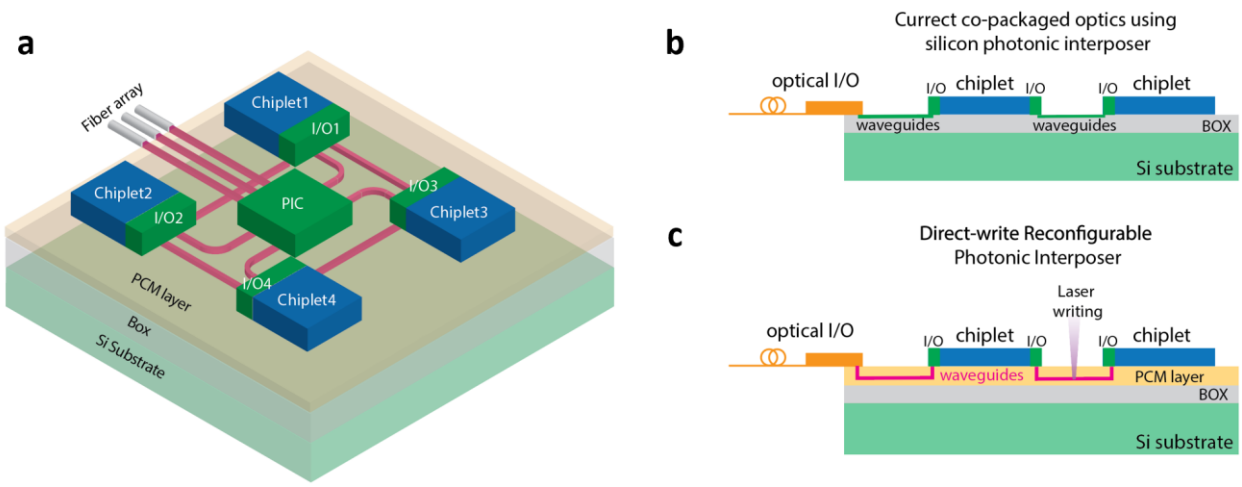

Figure 5.1 **a.** Schematic of the electronic ASIC chiplets connected within a phase-change material-based optical interposer for future SoC system. **b.** The currently proposed electronics/optics co-package design uses conventional silicon waveguides or fibers to connect chiplets. In this scheme, it brings challenges in alignment, reprogramming, and energy efficiency for the co-package of the SoC. **c.** The proposed phase-change material-based optical interposer. The interposer is nonvolatile and could be arbitrarily written and rewritten for various purposes.

Optical communications have been progressively supplanting electrical cabling in high-performance computing (HPC) and data center applications [10,55]. This transformation has replaced copper cables running between electrical faceplate ports with pluggable optical transceivers connected via fiber cables. However, as faceplates reach limitations concerning mechanical and thermal factors, there's a shift occurring in optical communications—moving from the faceplate to within the package itself.

The optical input/output (OIO) eliminates electrical I/O bottlenecks and transcends process limitations to unleash the next wave of innovation in semiconductor and data center design. As depicted in Figure 5.1, the paradigm of in-package optical I/O represents a groundbreaking approach. It involves the integration of application-specific integrated circuits (ASICs) chiplets with optical I/O modules within a multi-chip package (MCP). Figure 5.1a shows four chiplets are integrated on a photonic interposer which enables inter-chiplet communication through optical connections. Among the chiplets, PICs can also be integrated and connected with chiplets designed as photonic ASICs or control units. The currently proposed electronics/optics co-package design uses conventional silicon waveguides or fibers to connect chiplets. In this scheme, it brings

challenges in alignment, reprogramming, and energy efficiency for the co-package of the SoC. We show that the technology we developed in this thesis can work to solve these challenges. The optical computing platform using phase-change material demonstrated in Chapter 2 and Chapter 3 is suitable for future photonic PIC ASICs as it requires minimum or no external power to maintain its state once the data is programmed and stored in the phase-change memory unit on the PICs simultaneously. In Chapter 4, we also show the direct laser writing technique which enables connecting and reconnecting chiplets into an arbitrary configuration. As a post-processing approach, the direct laser writing of waveguides doesn't suffer from the alignment problems typically faced by the solutions at present. It is also worth noting that since the high optical confinement in the PCM waveguide, optical interconnections with higher density can be achieved, which would also potentially further increase the communication bandwidth in this SoC system.

BIBLIOGRAPHY

[1] W. Bogaerts, D. Pérez, J. Capmany, D. A. B. Miller, J. Poon, D. Englund, F. Morichetti, and A. Melloni, *Programmable Photonic Circuits*, Nature **586**, 207 (2020).

[2] D. Nikolova, S. Rumley, D. Calhoun, Q. Li, R. Hendry, P. Samadi, and K. Bergman, *Scaling Silicon Photonic Switch Fabrics for Data Center Interconnection Networks*, Opt Express **23**, 1159 (2015).

[3] Y. Li, Y. Zhang, L. Zhang, and A. W. Poon, *Silicon and Hybrid Silicon Photonic Devices for Intra-Datacenter Applications: State of the Art and Perspectives*, Photonics Res **3**, B10 (2015).

[4] K. Y. Yang, C. Shirpurkar, A. D. White, J. Zang, L. Chang, F. Ashtiani, M. A. Guidry, D. M. Lukin, S. V Pericherla, and J. Yang, *Multi-Dimensional Data Transmission Using Inverse-Designed Silicon Photonics and Microcombs*, Nat Commun **13**, 7862 (2022).

[5] 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 **7**, 7430 (2017).

[6] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, and D. Englund, *Deep Learning with Coherent Nanophotonic Circuits*, Nat Photonics **11**, 441 (2017).

[7] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, and A. S. Raja, *Parallel Convolutional Processing Using an Integrated Photonic Tensor Core*, Nature **589**, 52 (2021).

[8] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, *Photonics for Artificial Intelligence and Neuromorphic Computing*, Nat Photonics **15**, 102 (2021).

[9] C. Huang, S. Fujisawa, T. F. de Lima, A. N. Tait, E. C. Blow, Y. Tian, S. Bilodeau, A. Jha, F. Yaman, and H.-T. Peng, *A Silicon Photonic–Electronic Neural Network for Fibre Nonlinearity Compensation*, Nat Electron **4**, 837 (2021).

[10] N. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, and B. Towles, *Tpu v4: An Optically Reconfigurable Supercomputer for Machine*

Learning with Hardware Support for Embeddings, in *Proceedings of the 50th Annual International Symposium on Computer Architecture* (2023), pp. 1–14.

[11] J. Wang, D. Bonneau, M. Villa, J. W. Silverstone, R. Santagati, S. Miki, T. Yamashita, M. Fujiwara, M. Sasaki, and H. Terai, *Chip-to-Chip Quantum Photonic Interconnect by Path-Polarization Interconversion*, Optica **3**, 407 (2016).

[12] J. Wang, S. Paesani, Y. Ding, R. Santagati, P. Skrzypczyk, A. Salavrakos, J. Tura, R. Augusiak, L. Mančinska, and D. Bacco, *Multidimensional Quantum Entanglement with Large-Scale Integrated Optics*, Science (1979) **360**, 285 (2018).

[13] X. Qiang, X. Zhou, J. Wang, C. M. Wilkes, T. Loke, S. O'Gara, L. Kling, G. D. Marshall, R. Santagati, and T. C. Ralph, *Large-Scale Silicon Quantum Photonics Implementing Arbitrary Two-Qubit Processing*, Nat Photonics **12**, 534 (2018).

[14] D. Llewellyn, Y. Ding, I. I. Faruque, S. Paesani, D. Bacco, R. Santagati, Y.-J. Qian, Y. Li, Y.-F. Xiao, and M. Huber, *Chip-to-Chip Quantum Teleportation and Multi-Photon Entanglement in Silicon*, Nat Phys **16**, 148 (2020).

[15] J. M. Arrazola, V. Bergholm, K. Brádler, T. R. Bromley, M. J. Collins, I. Dhand, A. Fumagalli, T. Gerrits, A. Goussev, and L. G. Helt, *Quantum Circuits with Many Photons on a Programmable Nanophotonic Chip*, Nature **591**, 54 (2021).

[16] J. Sun, E. Timurdogan, A. Yaacobi, E. S. Hosseini, and M. R. Watts, *Large-Scale Nanophotonic Phased Array*, Nature **493**, 195 (2013).

[17] C. V. Poulton, M. J. Byrd, P. Russo, E. Timurdogan, M. Khandaker, D. Vermeulen, and M. R. Watts, *Long-Range LiDAR and Free-Space Data Communication with High-Performance Optical Phased Arrays*, IEEE Journal of Selected Topics in Quantum Electronics **25**, 1 (2019).

[18] J. Riemensberger, A. Lukashchuk, M. Karpov, W. Weng, E. Lucas, J. Liu, and T. J. Kippenberg, *Massively Parallel Coherent Laser Ranging Using a Soliton Microcomb*, Nature **581**, 164 (2020).

[19] X. Zhang, K. Kwon, J. Henriksson, J. Luo, and M. C. Wu, *A Large-Scale Microelectromechanical-Systems-Based Silicon Photonics LiDAR*, Nature **603**, 253 (2022).

[20] B. Li, Q. Lin, and M. Li, *Frequency–Angular Resolving LiDAR Using Chip-Scale Acousto-Optic Beam Steering*, Nature **620**, 316 (2023).

[21] A. Lukashchuk, J. Riemensberger, A. Tusnin, J. Liu, and T. J. Kippenberg, *Chaotic Microcomb-Based Parallel Ranging*, Nat Photonics **17**, 814 (2023).

[22] Q. Fang, J. F. Song, T.-Y. Liow, H. Cai, M. Bin Yu, G. Q. Lo, and D.-L. Kwong, *Ultralow Power Silicon Photonics Thermo-Optic Switch with Suspended Phase Arms*, IEEE Photonics Technology Letters **23**, 525 (2011).

[23] C. Joshi, J. K. Jang, K. Luke, X. Ji, S. A. Miller, A. Klenner, Y. Okawachi, M. Lipson, and A. L. Gaeta, *Thermally Controlled Comb Generation and Soliton Modelocking in Microresonators*, Opt Lett **41**, 2565 (2016).

[24] G. Liang, H. Huang, A. Mohanty, M. C. Shin, X. Ji, M. J. Carter, S. Shrestha, M. Lipson, and N. Yu, *Robust, Efficient, Micrometre-Scale Phase Modulators at Visible Wavelengths*, Nat Photonics **15**, 908 (2021).

[25] T. A. Ibrahim, W. Cao, Y. Kim, J. Li, J. Goldhar, P.-T. Ho, and C. H. Lee, *All-Optical Switching in a Laterally Coupled Microring Resonator by Carrier Injection*, IEEE Photonics Technology Letters **15**, 36 (2003).

[26] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, *Micrometre-Scale Silicon Electro-Optic Modulator*, Nature **435**, 325 (2005).

[27] S. Moazeni, S. Lin, M. Wade, L. Alloatti, R. J. Ram, M. Popović, and V. Stojanović, *A 40-Gb/s PAM-4 Transmitter Based on a Ring-Resonator Optical DAC in 45-Nm SOI CMOS*, IEEE J Solid-State Circuits **52**, 3503 (2017).

[28] A. H. Atabaki, S. Moazeni, F. Pavanello, H. Gevorgyan, J. Notaros, L. Alloatti, M. T. Wade, C. Sun, S. A. Kruger, and H. Meng, *Integrating Photonics with Silicon Nanoelectronics for the next Generation of Systems on a Chip*, Nature **556**, 349 (2018).

[29] C. T. Phare, Y.-H. Daniel Lee, J. Cardenas, and M. Lipson, *Graphene Electro-Optic Modulator with 30 GHz Bandwidth*, Nat Photonics **9**, 511 (2015).

[30] C. Wang, M. Zhang, B. Stern, M. Lipson, and M. Lončar, *Nanophotonic Lithium Niobate Electro-Optic Modulators*, Opt Express **26**, 1547 (2018).

[31] C. Wang, M. Zhang, X. Chen, M. Bertrand, A. Shams-Ansari, S. Chandrasekhar, P. Winzer, and M. Lončar, *Integrated Lithium Niobate Electro-Optic Modulators Operating at CMOS-Compatible Voltages*, Nature **562**, 101 (2018).

[32] H. Sattari, T. Graziosi, M. Kiss, T. J. Seok, S. Han, M. C. Wu, and N. Quack, *Silicon Photonic MEMS Phase-Shifter*, Opt Express **27**, 18959 (2019).

[33] P. Edinger, A. Y. Takabayashi, C. Errando-Herranz, U. Khan, H. Sattari, P. Verheyen, W. Bogaerts, N. Quack, and K. B. Gylfason, *Silicon Photonic Microelectromechanical Phase Shifters for Scalable Programmable Photonics*, Opt Lett **46**, 5671 (2021).

[34] D. Lencer, M. Salinga, B. Grabowski, T. Hickel, J. Neugebauer, and M. Wuttig, *A Map for Phase-Change Materials*, Nat Mater **7**, 972 (2008).

[35] C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, *Integrated All-Photonic Non-Volatile Multi-Level Memory*, Nat Photonics **9**, 725 (2015).

[36] R. Chen, Z. Fang, C. Perez, F. Miller, K. Kumari, A. Saxena, J. Zheng, S. J. Geiger, K. E. Goodson, and A. Majumdar, *Non-Volatile Electrically Programmable Integrated Photonics with a 5-Bit Operation*, Nat Commun **14**, 3465 (2023).

[37] Y. Zhang, J. B. Chou, J. Li, H. Li, Q. Du, A. Yadav, S. Zhou, M. Y. Shalaginov, Z. Fang, and H. Zhong, *Broadband Transparent Optical Phase Change Materials for High-Performance Nonvolatile Photonics*, Nat Commun **10**, 4279 (2019).

[38] C. Rios, M. Stegmaier, Z. Cheng, N. Youngblood, C. D. Wright, W. H. P. Pernice, and H. Bhaskaran, *Controlled Switching of Phase-Change Materials by Evanescent-Field Coupling in Integrated Photonics*, Opt Mater Express **8**, 2455 (2018).

[39] N. Farmakidis, N. Youngblood, X. Li, J. Tan, J. L. Swett, Z. Cheng, C. D. Wright, W. H. P. Pernice, and H. Bhaskaran, *Plasmonic Nanogap Enhanced Phase-Change Devices with Dual Electrical-Optical Functionality*, Sci Adv **5**, eaaw2687 (2019).

[40] X. Li, N. Youngblood, C. Ríos, Z. Cheng, C. D. Wright, W. H. P. Pernice, and H. Bhaskaran, *Fast and Reliable Storage Using a 5 Bit, Nonvolatile Photonic Memory Cell*, Optica **6**, 1 (2019).

[41] M. Wuttig, H. Bhaskaran, and T. Taubner, *Phase-Change Materials for Non-Volatile Photonic Applications*, Nat Photonics **11**, 465 (2017).

[42] M. Wuttig and N. Yamada, *Phase-Change Materials for Rewriteable Data Storage*, Nat Mater **6**, 824 (2007).

[43] H. Zhang, L. Zhou, B. M. A. Rahman, X. Wu, L. Lu, Y. Xu, J. Xu, J. Song, Z. Hu, and L. Xu, *Ultracompact Si-GST Hybrid Waveguides for Nonvolatile Light Wave Manipulation*, IEEE Photonics J **10**, 1 (2017).

[44] C. Wu, H. Yu, H. Li, X. Zhang, I. Takeuchi, and M. Li, *Low-Loss Integrated Photonic Switch Using Subwavelength Patterned Phase Change Material*, ACS Photonics **6**, 87 (2018).

[45] M. Delaney, I. Zeimpekis, D. Lawson, D. W. Hewak, and O. L. Muskens, *A New Family of Ultralow Loss Reversible Phase-change Materials for Photonic Integrated Circuits: Sb2S3 and Sb2Se3*, Adv Funct Mater **30**, 2002447 (2020).

[46] M. Delaney, I. Zeimpekis, H. Du, X. Yan, M. Banakar, D. J. Thomson, D. W. Hewak, and O. L. Muskens, *Nonvolatile Programmable Silicon Photonics Using an Ultralow-Loss Sb2Se3 Phase Change Material*, Sci Adv **7**, eabg3500 (2021).

[47] Z. Fang, B. Mills, R. Chen, J. Zhang, P. Xu, J. Hu, and A. Majumdar, *Arbitrary Programming of Racetrack Resonators Using Low-Loss Phase-Change Material Sb2Se3*, Nano Lett **24**, 97 (2023).

[48] P. Xu, J. Zheng, J. K. Doylend, and A. Majumdar, *Low-Loss and Broadband Nonvolatile Phase-Change Directional Coupler Switches*, ACS Photonics **6**, 553 (2019).

[49] H. Zhang, L. Zhou, L. Lu, J. Xu, N. Wang, H. Hu, B. M. A. Rahman, Z. Zhou, and J. Chen, *Miniature Multilevel Optical Memristive Switch Using Phase Change Material*, ACS Photonics **6**, 2205 (2019).

[50] J. Zheng, Z. Fang, C. Wu, S. Zhu, P. Xu, J. K. Doylend, S. Deshmukh, E. Pop, S. Dunham, and M. Li, *Nonvolatile Electrically Reconfigurable Integrated Photonic Switch Enabled by a Silicon PIN Diode Heater*, Advanced Materials **32**, 2001218 (2020).

[51] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, *All-Optical Spiking Neurosynaptic Networks with Self-Learning Capabilities*, Nature **569**, 208 (2019).

[52] Z. Cheng, C. Ríos, W. H. P. Pernice, C. D. Wright, and H. Bhaskaran, *On-Chip Photonic Synapse*, Sci Adv **3**, e1700160 (2017).

[53] Z. Fang, R. Chen, J. Zheng, A. I. Khan, K. M. Neilson, S. J. Geiger, D. M. Callahan, M. G. Moebius, A. Saxena, and M. E. Chen, *Ultra-Low-Energy Programmable Non-Volatile Silicon Photonics Based on Phase-Change Materials with Graphene Heaters*, Nat Nanotechnol **17**, 842 (2022).

[54] R. Athale and D. Psaltis, *Optical Computing: Past and Future*, Opt Photonics News **27**, 32 (2016).

[55] P. L. McMahon, *The Physics of Optical Computing*, Nature Reviews Physics **5**, 717 (2023).

[56] N. Jones, *How to Stop Data Centres from Gobbling up the World's Electricity*, Nature **561**, 163 (2018).

[57] Y. Shen, X. Meng, Q. Cheng, S. Rumley, N. Abrams, A. Gazman, E. Manzhosov, M. S. Glick, and K. Bergman, *Silicon Photonics for Extreme Scale Systems*, Journal of Lightwave Technology **37**, 245 (2019).

[58] D. Pérez-López, A. López, P. DasMahapatra, and J. Capmany, *Multipurpose Self-Configuration of Programmable Photonic Circuits*, Nat Commun **11**, 6359 (2020).

[59] C. H. Chu, M. L. Tseng, J. Chen, P. C. Wu, Y. Chen, H. Wang, T. Chen, W. T. Hsieh, H. J. Wu, and G. Sun, *Active Dielectric Metasurface Based on Phase-change Medium*, Laser Photon Rev **10**, 986 (2016).

[60] X. Yin, T. Steinle, L. Huang, T. Taubner, M. Wuttig, T. Zentgraf, and H. Giessen, *Beam Switching and Bifocal Zoom Lensing Using Active Plasmonic Metasurfaces*, Light Sci Appl **6**, e17016 (2017).

[61] Z. Cheng, C. Ríos, N. Youngblood, C. D. Wright, W. H. P. Pernice, and H. Bhaskaran, *Device-level Photonic Memories and Logic Applications Using Phase-change Materials*, Advanced Materials **30**, 1802435 (2018).

[62] I. Chakraborty, G. Saha, and K. Roy, *Photonic In-Memory Computing Primitive for Spiking Neural Networks Using Phase-Change Materials*, Phys Rev Appl **11**, 014063 (2019).

[63] C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, *Programmable Phase-Change Metasurfaces on Waveguides for Multimode Photonic Convolutional Neural Network*, Nat Commun **12**, 96 (2021).

[64] Z. Li, M.-H. Kim, C. Wang, Z. Han, S. Shrestha, A. C. Overvig, M. Lu, A. Stein, A. M. Agarwal, and M. Lončar, *Controlling Propagation and Coupling of Waveguide Modes Using Phase-Gradient Metasurfaces*, Nat Nanotechnol **12**, 675 (2017).

[65] J.-W. Park, S. H. Eom, H. Lee, J. L. F. Da Silva, Y.-S. Kang, T.-Y. Lee, and Y. H. Khang, *Optical Properties of Pseudobinary GeTe, Ge 2 Sb 2 Te 5, GeSb 2 Te 4, GeSb 4 Te 7, and Sb 2 Te 3 from Ellipsometry and Density Functional Theory*, Phys Rev B **80**, 115209 (2009).

[66] M. Le Gallo, A. Sebastian, R. Mathis, M. Manica, H. Giefers, T. Tuma, C. Bekas, A. Curioni, and E. Eleftheriou, *Mixed-Precision in-Memory Computing*, Nat Electron **1**, 246 (2018).

[67] M. A. Nahmias, T. F. De Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, *Photonic Multiply-Accumulate Operations for Neural Networks*, IEEE Journal of Selected Topics in Quantum Electronics **26**, 1 (2019).

[68] B. Marr, B. Degnan, P. Hasler, and D. Anderson, *Scaling Energy per Operation via an Asynchronous Pipeline*, IEEE Trans Very Large Scale Integr VLSI Syst **21**, 147 (2012).

[69] M. M. Waldrop, *The Chips Are down for Moore's Law*, Nature News **530**, 144 (2016).

[70] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi, *Scaling for Edge Inference of Deep Neural Networks*, Nat Electron **1**, 216 (2018).

[71] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, *All-Optical Machine Learning Using Diffractive Deep Neural Networks*, Science (1979) **361**, 1004 (2018).

[72] 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 **588**, 39 (2020).

[73] T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, *Large-Scale Neuromorphic Optoelectronic Computing with a Reconfigurable Diffractive Processing Unit*, Nat Photonics **15**, 367 (2021).

[74] L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, *Freely Scalable and Reconfigurable Optical Hardware for Deep Learning*, Sci Rep **11**, 3144 (2021).

[75] A. F. Murray and P. J. Edwards, *Enhanced MLP Performance and Fault Tolerance Resulting from Synaptic Weight Noise during Training*, IEEE Trans Neural Netw **5**, 792 (1994).

[76] S. Moon, K. Shin, and D. Jeon, *Enhancing Reliability of Analog Neural Network Processors*, IEEE Trans Very Large Scale Integr VLSI Syst **27**, 1455 (2019).

[77] V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, *Accurate Deep Neural Network Inference Using Computational Phase-Change Memory*, Nat Commun **11**, 2473 (2020).

[78] C. M. Bishop, *Training with Noise Is Equivalent to Tikhonov Regularization*, Neural Comput **7**, 108 (1995).

[79] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, *Weight Uncertainty in Neural Network*, in *International Conference on Machine Learning* (PMLR, 2015), pp. 1613–1622.

[80] A. S. Rekhi, B. Zimmer, N. Nedovic, N. Liu, R. Venkatesan, M. Wang, B. Khailany, W. J. Dally, and C. T. Gray, *Analog/Mixed-Signal Hardware Error Modeling for Deep Learning Inference*, in *Proceedings of the 56th Annual Design Automation Conference 2019* (2019), pp. 1–6.

[81] M. Klachko, M. R. Mahmoodi, and D. Strukov, *Improving Noise Tolerance of Mixed-Signal Neural Networks*, in *2019 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2019), pp. 1–8.

[82] A. Sebastian, T. Tuma, N. Papandreou, M. Le Gallo, L. Kull, T. Parnell, and E. Eleftheriou, *Temporal Correlation Detection Using Computational Phase-Change Memory*, Nat Commun **8**, 1115 (2017).

[83] C. Ríos, N. Youngblood, Z. Cheng, M. Le Gallo, W. H. P. Pernice, C. D. Wright, A. Sebastian, and H. Bhaskaran, *In-Memory Computing on a Photonic Platform*, Sci Adv **5**, eaau5759 (2019).

[84] L. G. Wright, T. Onodera, M. M. Stein, T. Wang, D. T. Schachter, Z. Hu, and P. L. McMahon, *Deep Physical Neural Networks Trained with Backpropagation*, Nature **601**, 549 (2022).

[85] D. Mengu, Y. Zhao, N. T. Yardimci, Y. Rivenson, M. Jarrahi, and A. Ozcan, *Misalignment Resilient Diffractive Optical Networks*, Nanophotonics **9**, 4207 (2020).

[86] D. Mengu, Y. Rivenson, and A. Ozcan, *Scale-, Shift-, and Rotation-Invariant Diffractive Optical Networks*, ACS Photonics **8**, 324 (2020).

[87] P. Lalanne, J.-C. Rodier, P. H. Chavel, E. Belhaire, and P. F. Garda, *Optoelectronic Devices for Boltzmann Machines and Simulated Annealing*, Optical Engineering **32**, 1904 (1993).

[88] A. Dupret, E. Belhaire, J.-C. Rodier, P. Lalanne, D. Prévost, P. Garda, and P. Chavel, *An Optoelectronic CMOS Circuit Implementing a Simulated Annealing Algorithm*, IEEE J Solid-State Circuits **31**, 1046 (1996).

[89] F. Cai, S. Kumar, T. Van Vaerenbergh, X. Sheng, R. Liu, C. Li, Z. Liu, M. Foltin, S. Yu, and Q. Xia, *Power-Efficient Combinatorial Optimization Using Intrinsic Noise in Memristor Hopfield Neural Networks*, Nat Electron **3**, 409 (2020).

[90] T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz, and E. Vianello, *In Situ Learning Using Intrinsic Memristor Variability via Markov Chain Monte Carlo Sampling*, Nat Electron **4**, 151 (2021).

[91] J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, *Generative or Discriminative? Getting the Best of Both Worlds*, Bayesian Statistics **8**, 3 (2007).

[92] T. Jebara, *Machine Learning: Discriminative and Generative*, Vol. 755 (Springer Science & Business Media, 2012).

[93] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski, *Deep Generative Stochastic Networks Trainable by Backprop*, in *International Conference on Machine Learning* (PMLR, 2014), pp. 226–234.

[94] C. R. S. Williams, J. C. Salevan, X. Li, R. Roy, and T. E. Murphy, *Fast Physical Random Number Generator Using Amplified Spontaneous Emission*, Opt Express **18**, 23584 (2010).

[95] X. Li, A. B. Cohen, T. E. Murphy, and R. Roy, *Scalable Parallel Physical Random Number Generator Based on a Superluminescent LED*, Opt Lett **36**, 1020 (2011).

[96] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets*, Adv Neural Inf Process Syst **27**, (2014).

[97] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, *Improved Techniques for Training Gans*, Adv Neural Inf Process Syst **29**, (2016).

[98] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, *Gans Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*, Adv Neural Inf Process Syst **30**, (2017).

[99] D. M. Baney, P. Gallion, and R. S. Tucker, *Theory and Measurement Techniques for the Noise Figure of Optical Amplifiers*, Optical Fiber Technology **6**, 122 (2000).

[100] G. Duan and E. Georgiev, *Nonwhite Photodetection Noise at the Ouput of an Optical Amplifier: Theory and Experiment*, in *Physics and Simulation of Optoelectronic Devices IX*, Vol. 4283 (SPIE, 2001), pp. 540–550.

[101] Y. Lin, H. Wu, B. Gao, P. Yao, W. Wu, Q. Zhang, X. Zhang, X. Li, F. Li, and J. Lu, *Demonstration of Generative Adversarial Network by Intrinsic Random Noises of Analog Rram Devices*, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), pp. 3–4.

[102] O. Krestinskaya, B. Choubey, and A. P. James, *Memristive GAN in Analog*, Sci Rep **10**, 5838 (2020).

[103] C. G. Turhan and H. S. Bilge, *Recent Trends in Deep Generative Models: A Review*, in *2018 3rd International Conference on Computer Science and Engineering (UBMK)* (IEEE, 2018), pp. 574–579.

[104] X.-Y. Xu, X.-L. Huang, Z.-M. Li, J. Gao, Z.-Q. Jiao, Y. Wang, R.-J. Ren, H. P. Zhang, and X.-M. Jin, *A Scalable Photonic Computer Solving the Subset Sum Problem*, Sci Adv **6**, eaay5853 (2020).

[105] 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*, Nat Commun **11**, 2154 (2020).

[106] F. Ashtiani, A. J. Geers, and F. Aflatouni, *An On-Chip Photonic Deep Neural Network for Image Classification*, Nature **606**, 501 (2022).

[107] X. Xu, G. Ren, T. Feleppa, X. Liu, A. Boes, A. Mitchell, and A. J. Lowery, *Self-Calibrating Programmable Photonic Integrated Circuits*, Nat Photonics **16**, 595 (2022).

[108] A. Z. Subramanian, E. Ryckeboer, A. Dhakal, F. Peyskens, A. Malik, B. Kuyken, H. Zhao, S. Pathak, A. Ruocco, and A. De Groote, *Silicon and Silicon Nitride Photonic Circuits for Spectroscopic Sensing On-a-Chip*, Photonics Res **3**, B47 (2015).

[109] E. Luan, H. Shoman, D. M. Ratner, K. C. Cheung, and L. Chrostowski, *Silicon Photonic Biosensors Using Label-Free Detection*, Sensors **18**, 3519 (2018).

[110] J. Liu, G. Huang, R. N. Wang, J. He, A. S. Raja, T. Liu, N. J. Engelsen, and T. J. Kippenberg, *High-Yield, Wafer-Scale Fabrication of Ultralow-Loss, Dispersion-Engineered Silicon Nitride Photonic Circuits*, Nat Commun **12**, 2236 (2021).

[111] R. Nagarajan, M. Kato, J. Pleumeekers, P. Evans, S. Corzine, S. Hurtt, A. Dentai, S. Murthy, M. Missey, and R. Muthiah, *InP Photonic Integrated Circuits*, IEEE Journal of Selected Topics in Quantum Electronics **16**, 1113 (2010).

[112] L. M. Augustin, R. Santos, E. den Haan, S. Kleijn, P. J. A. Thijs, S. Latkowski, D. Zhao, W. Yao, J. Bolk, and H. Ambrosius, *InP-Based Generic Foundry Platform for Photonic Integrated Circuits*, IEEE Journal of Selected Topics in Quantum Electronics **24**, 1 (2017).

[113] L. S. Madsen, F. Laudenbach, M. F. Askarani, F. Rortais, T. Vincent, J. F. F. Bulmer, F. M. Miatto, L. Neuhaus, L. G. Helt, and M. J. Collins, *Quantum Computational Advantage with a Programmable Photonic Processor*, Nature **606**, 75 (2022).

[114] W. Liu, M. Li, R. S. Guzzon, E. J. Norberg, J. S. Parker, M. Lu, L. A. Coldren, and J. Yao, *A Fully Reconfigurable Photonic Integrated Signal Processor*, Nat Photonics **10**, 190 (2016).

[115] W. Bogaerts, M. Fiers, and P. Dumon, *Design Challenges in Silicon Photonics*, IEEE Journal of Selected Topics in Quantum Electronics **20**, 1 (2013).

[116] Z. Lu, J. Jhoja, J. Klein, X. Wang, A. Liu, J. Flueckiger, J. Pond, and L. Chrostowski, *Performance Prediction for Silicon Photonics Integrated Circuits with Layout-Dependent Correlated Manufacturing Variability*, Opt Express **25**, 9712 (2017).

[117] W. Bogaerts and L. Chrostowski, *Silicon Photonics Circuit Design: Methods, Tools and Challenges*, Laser Photon Rev **12**, 1700237 (2018).

[118] C. Wu, H. Deng, Y.-S. Huang, H. Yu, I. Takeuchi, C. A. Ríos Ocampo, and M. Li, *Freeform Direct-Write and Rewritable Photonic Integrated Circuits in Phase-Change Thin Films*, Sci Adv **10**, eadk1361 (2024).

[119] B. Gholipour, J. Zhang, K. F. MacDonald, D. W. Hewak, and N. I. Zheludev, *An All-optical, Non-volatile, Bidirectional, Phase-change Meta-switch*, Advanced Materials **25**, 3050 (2013).

[120] Q. Wang, E. T. F. Rogers, B. Gholipour, C.-M. Wang, G. Yuan, J. Teng, and N. I. Zheludev, *Optically Reconfigurable Metasurfaces and Photonic Devices Based on Phase Change Materials*, Nat Photonics **10**, 60 (2016).

[121] H. Liu, W. Dong, H. Wang, L. Lu, Q. Ruan, Y. S. Tan, R. E. Simpson, and J. K. W. Yang, *Rewritable Color Nanoprints in Antimony Trisulfide Films*, Sci Adv **6**, eabb7171 (2020).

[122] K. Chaudhary, M. Tamagnone, X. Yin, C. M. Spägele, S. L. Oscurato, J. Li, C. Persch, R. Li, N. A. Rubin, and L. A. Jauregui, *Polariton Nanophotonics Using Phase-Change Materials*, Nat Commun **10**, 4487 (2019).

[123] M. Deubel, G. Von Freymann, M. Wegener, S. Pereira, K. Busch, and C. M. Soukoulis, *Direct Laser Writing of Three-Dimensional Photonic-Crystal Templates for Telecommunications*, Nat Mater **3**, 444 (2004).

[124] A. Politi, M. J. Cryan, J. G. Rarity, S. Yu, and J. L. O'brien, *Silica-on-Silicon Waveguide Quantum Circuits*, Science (1979) **320**, 646 (2008).

[125] G. D. Marshall, A. Politi, J. C. F. Matthews, P. Dekker, M. Ams, M. J. Withford, and J. L. O'Brien, *Laser Written Waveguide Photonic Quantum Circuits*, Opt Express **17**, 12546 (2009).

[126] T. Meany, M. Gräfe, R. Heilmann, A. Perez-Leija, S. Gross, M. J. Steel, M. J. Withford, and A. Szameit, *Laser Written Circuits for Quantum Photonics*, Laser Photon Rev **9**, 363 (2015).

[127] O. M. Efimov, L. B. Glebov, K. A. Richardson, E. Van Stryland, T. Cardinal, S. H. Park, M. Couzi, and J. L. Bruneel, *Waveguide Writing in Chalcogenide Glasses by a Train of Femtosecond Laser Pulses*, Opt Mater (Amst) **17**, 379 (2001).

[128] A. Zoubir, M. Richardson, C. Rivero, A. Schulte, C. Lopez, K. Richardson, N. Hô, and R. Vallée, *Direct Femtosecond Laser Writing of Waveguides in As 2 S 3 Thin Films*, Opt Lett **29**, 748 (2004).

[129] M. A. Hughes, W. Yang, and D. W. Hewak, *Spectral Broadening in Femtosecond Laser Written Waveguides in Chalcogenide Glass*, JOSA B **26**, 1370 (2009).

[130] B. J. Eggleton, B. Luther-Davies, and K. Richardson, *Chalcogenide Photonics*, Nat Photonics **5**, 141 (2011).

[131] T. Wu, M. Menarini, Z. Gao, and L. Feng, *Lithography-Free Reconfigurable Integrated Photonic Processor*, Nat Photonics 1 (2023).

[132] L. Zhuang, C. G. H. Roeloffzen, M. Hoekman, K.-J. Boller, and A. J. Lowery, *Programmable Photonic Signal Processor Chip for Radiofrequency Applications*, Optica **2**, 854 (2015).

[133] O. Daulay, G. Liu, K. Ye, R. Botter, Y. Klaver, Q. Tan, H. Yu, M. Hoekman, E. Klein, and C. Roeloffzen, *Ultrahigh Dynamic Range and Low Noise Figure Programmable Integrated Microwave Photonic Filter*, Nat Commun **13**, 7798 (2022).

[134] R. A. Cohen, O. Amrani, and S. Ruschin, *Response Shaping with a Silicon Ring Resonator via Double Injection*, Nat Photonics **12**, 706 (2018).

[135] S. Pai, Z. Sun, T. W. Hughes, T. Park, B. Bartlett, I. A. D. Williamson, M. Minkov, M. Milanizadeh, N. Abebe, and F. Morichetti, *Experimentally Realized in Situ Backpropagation for Deep Learning in Photonic Neural Networks*, Science (1979) **380**, 398 (2023).

VITA

Changming was born in Hefei, China which is located in eastern Anhui province. He received a bachelor's degree in Applied Physics from the University of Science and Technology of China in 2014. From 2015 to 2017, Changming spent three years at the Hong Kong University of Science and Technology and received a master’s degree in Physics. He then started the journey of the Ph.D. in Electrical Engineering at the University of Minnesota under the supervision of Prof. Mo Li and then transferred with the Li group to the University of Washington in 2018. He has developed broad interests in programable photonics based on optical phase-change material and optoelectronics during his Ph.D.