Research Article Vol. 12, No. 10 / October 2024 / Photonics Research 2279

PHOTONICS Research

Mode-multiplexed photonic integrated vector dot-product core from inverse design

Zheyuan Zhu,[1,*] Raktim Sarma,[2] Seth Smith-Dryden,[1] Guifang Li,[1] and Shuo S. Pang[1]

[1]CREOL, The College of Optics and Photonics, University of Central Florida, Orlando, Florida 32816-2700, USA
[2]Center for Integrated Nanotechnologies, Sandia National Laboratories, Albuquerque, New Mexico 87123, USA
*Corresponding author: zheyuan.zhu@ucf.edu

Received 3 April 2024; revised 24 June 2024; accepted 22 July 2024; posted 23 July 2024 (Doc. ID 524419); published 1 October 2024

Photonic computing has the potential to harness the full degrees of freedom (DOFs) of the light field, including the wavelength, spatial mode, spatial location, phase quadrature, and polarization, to achieve a higher level of computing parallelism and scalability than digital electronic processors. While multiplexing using the wavelength and other DOFs can be readily integrated on silicon photonics platforms with compact footprints, conventional mode-division multiplexed (MDM) photonic designs occupy areas exceeding tens to hundreds of microns for a few spatial modes, significantly limiting their scalability. Here, we utilize inverse design to demonstrate an ultra-compact photonic computing core that calculates vector dot products based on MDM coherent mixing. Our dot-product core integrates the functionalities of two-mode multiplexers and one multimode coherent mixer within a nominal footprint of 5 μm × 3 μm. We have experimentally demonstrated computing examples on the fabricated dot-product core, including complex number multiplication and motion estimation using optical flow. The compact dot-product core design enables large-scale on-chip integration in a parallel photonic computing primitive cluster for high-throughput scientific computing and computer vision tasks.

https://doi.org/10.1364/PRJ.524419

1. INTRODUCTION

Vector, matrix, or tensor calculations are the fundamental building blocks of modern scientific computing. The underlying core components of these computing tasks are basic linear algebra subprograms (BLASs) that provide hardware implementation of the arithmetic operations between vectors (level 1), vector and matrix (level 2), and matrices (level 3), each building upon the previous level [1]. Given its unique role as a BLAS level 1 routine, efficient and scalable vector dot-product calculation is crucial to achieving optimal performances in more complex and computationally intensive operations. In traditional uniprocessor digital computers, the central processing unit (CPU) executes a single basic operation, such as addition, multiplication, or fused multiply–add (FMA), on a single data stream, a process known as single instruction stream, single data stream (SISD), as shown in Fig. 1(a). The sequential execution and repeated data access of SISD compromise the computation speed and efficiency in vector- and matrix-based operations. Single instruction stream, multiple data streams (SIMD), as shown in Fig. 1(b), which simultaneously applies an arithmetic operation to multiple data streams [2], has been adopted in virtually all modern CPUs and stream processors in GPUs. These processors incorporate dedicated SIMD tiles of cascaded FMA units with pipelined inputs to accelerate vector instructions [3]. Because caching the intermediate results is still necessary to ensure timing closure in electronics, the computing throughput per unit area is usually on the order of 0.1 tera operations per second per millimeter square (TOPS/mm^2), and the vector length is typically limited to several hundred, even with a highly optimized layout of logic and memory units within an SIMD engine [3,4].

Recently, driven by the computing demand in the artificial intelligence (AI), analog computing platforms based on integrated photonic devices [5,6] have demonstrated the potential of higher efficiency and computing throughput than the electronic counterparts, due to the intrinsically passive photonic multiply–accumulate (MAC) operations without intermediate memory access [7]. Figure 1(c) illustrates a photonic computing design based on two coherent mixers without parallelization in DOF of light, much like the SISD architecture in digital computing. In a single coherent mixing unit, the two inputs of electrical fields encode the numbers a and b in their amplitudes. After splitting and balanced detection, the output is proportional to their product $\mathrm{Re}\{a^* b\}$ [8]. To perform dot products between two N-element vectors, N sets of mixers and balanced photodiodes are required, and the intermediate element-wise products must first be individually digitized and then summed in the post-processing stage. Due to the power consumption of analog-to-digital converters (ADCs) [9] required in the design, coherent mixing without data-level

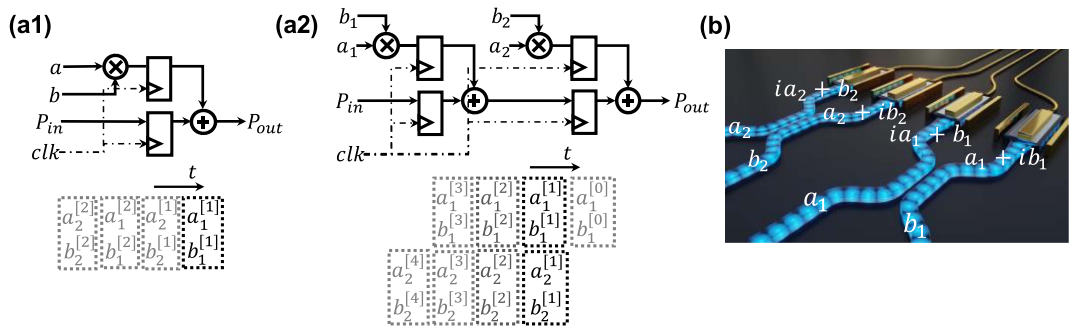

Fig. 1. (a) Electronic and (b) photonic implementations of SISD and SIMD operations. (a1) An SISD electronic arithmetic unit that performs multiplication and addition. (a2) SIMD design with multiple, pipelined inputs for dot-product calculation. (b) Individual single-mode coherent mixers as multiplier units without parallelism, equivalent to SISD architecture in digital electronics.

parallelism suffers from low efficiency when handling large vectors.

Similar to the transition from SISD to SIMD architecture in digital processors, using wavelength- or mode-division multiplexed (WDM or MDM) photonic signals enables a single coherent detection unit, consisting of a 2 × 2 coherent mixer and a pair of balanced photodiodes, to simultaneously process multiple data inputs in parallel [10,11]. Leveraging the intrinsic orthogonality of the light fields, coherent photonic MAC operations with multiplexed signals naturally accumulate the intermediate elementwise products between two vectors, and thus could achieve two- or threefold lower power consumption than the nonmultiplexed designs [12]. While WDM-based photonic processing devices have matured into practice to some extent in AI-related computing applications [13,14], MDM-based devices only begin to emerge as a viable approach in high-bandwidth optical communication [15,16], and their applications in parallel photonic computing are yet to be exploited.

Although utilizing MDM can lead to significant advances in high-bandwidth optical communication and photonic computing, a major bottleneck to high density integration is the large footprint usually associated with these MDM-based nanophotonic devices. In addition, different from the conventional MDM components used for optical communication, the complexity of photonic computing often necessitates two or more traditional MDM building blocks to implement the arithmetic operations. In this work, we present an end-to-end MDM-based photonic design that integrates the functionalities of multiple MDM blocks, resulting in a compact footprint for vector and/or matrix-based SIMD computing applications. Combined with peripheral electronics and algorithms targeting our photonic platform, we have experimentally demonstrated vector-dot product, complex number multiplication, and a computer vision task on a fabricated MDM-based photonic dot-product core.

2. PRINCIPLE OF OPERATION

A. MDM Coherent Photonic Dot-Product Core

Figure 2(a) shows an implementation of photonic dot-product core based on conventional MDM components in optical communications. The elements in the vector, a_1 (b_1) and a_2 (b_2), are mapped to the electric field profiles of the fundamental (ψ_{I}, TE0) and the second order (ψ_{II}, TE1) TE modes of a few-mode waveguide via mode multiplexers (MUXs). The mode-multiplexed photonic signals, $E_a = a_1\psi_{\mathrm{I}} + a_2\psi_{\mathrm{II}}$ and $E_b = b_1\psi_{\mathrm{I}} + b_2\psi_{\mathrm{II}}$, undergo coherent mixing via multimode interference (MMI), producing the electrical fields on the upper and lower arms $E_p = \frac{1}{\sqrt{2}}(E_a + iE_b)$ and $E_n = \frac{1}{\sqrt{2}}(iE_a + E_b)$. Based on the orthogonality between ψ_{I} and ψ_{II}, the difference between the overall intensity of the upper and lower outputs $I_{\mathrm{diff}} = |E_p|^2 - |E_n|^2$ produces the dot-product between vectors $\vec{a}$ and $\vec{b}$. The functionality of the conventional MDM dot-product design can be expressed as a Kronecker product (denoted by $\otimes$) between a 3 dB coupling matrix representing the MMI, and an identity matrix representing the MUX, as in

$$\begin{bmatrix} E_{\mathrm{II}p} \\ E_{\mathrm{I}p} \\ E_{\mathrm{I}n} \\ E_{\mathrm{II}n} \end{bmatrix} = \frac{1}{\sqrt{2}}\left(\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} a_1 \\ a_2 \\ b_2 \\ b_1 \end{bmatrix}. \quad (1)$$

Using conventional MDM components, the dot-product core requires two MUXs and one MMI. Each MUX occupies at least 20 μm × 4 μm in footprint [17,18], which is required by the adiabatic taper. For a 2 × 2 MMI, a footprint of 40 μm × 6 μm [19,20] is required to match the first Talbot distance. The overall footprint of a conventional dot-product core is thus larger than 50 μm × 10 μm.

Figure 2(b) shows our topologically optimized mode-multiplexed photonic vector dot-product core that integrates the functionalities of two MUXs and one MMI within a 5 μm × 3 μm footprint. Compared to the behavior of the electrical field inside a conventional multimode photonic design [Fig. 2(b1)], in which the regions for mode multiplexing and mixing are clearly distinguishable, the integrated dot-product core does not perform an intermediate conversion step of the input electrical fields onto the spatial mode basis. The end-to-end transformation of the electrical field by the integrated core, expressed as matrix S_i in

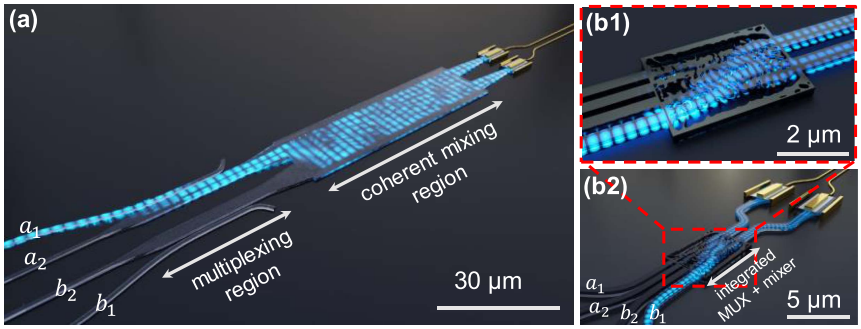

Fig. 2. Photonic implementation of vector dot-product core based on mode-division multiplexing. (a) Implementation based on the traditional MDM infrastructure in optical communications, using two MUXs and one MMI. (b) An end-to-end photonic dot-product core that integrates the functionalities of two MUXs and one MMI. The inset shows the photonic structure from inverse design.

$$\begin{bmatrix} E_{\mathrm{II}p} \\ E_{\mathrm{I}p} \\ E_{\mathrm{I}n} \\ E_{\mathrm{II}n} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 0 & i \\ 0 & 1 & i & 0 \\ 0 & i & 1 & 0 \\ i & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ b_2 \\ b_1 \end{bmatrix}, \tag{2}$$

contributes to its compact footprint.

The ultracompact footprint addresses one of the fundamental bottlenecks for utilizing MDM-based approaches for photonic computing and paves the way for high-density integration of the core in a parallel computing array.

B. End-to-End Design of Photonic Dot-Product Core

The photonic core was inversely designed on a silicon-on-insulator (SOI) platform by optimizing the structure that maximizes the coupling efficiency from the inputs into the target electric field profiles. The design process follows a gradient-based paradigm that tunes the distribution of the relative permittivity $\boldsymbol{\varepsilon}_r$ on the silicon layer as the design parameters [21–24]. The parameters are updated along the gradient direction of the objective function

$$l = \sum_{j=1}^{J} |\boldsymbol{E}_{t_j}^{\dagger} \boldsymbol{E}_j(\boldsymbol{\varepsilon}_r)|^2. \tag{3}$$

Here, l calculates the overlap integral between the target field $\boldsymbol{E}_{t_j}$ at the output location and the field $\boldsymbol{E}_j$ within the structure, $\boldsymbol{\varepsilon}_r$ is the three-dimensional distribution of relative permittivity, and $(\cdot)^{\dagger}$ denotes the matrix conjugate transpose. We set the fundamental or second-order TE eigenmodes in the few-mode output waveguides as the target fields $\boldsymbol{E}_{t_j}$. The summation over j aggregates the contributions from all four pairs of output and target fields. The field $\boldsymbol{E}_j$ in the device satisfies the finite-difference frequency domain (FDFD) Maxwell equations in matrix form, expressed as

$$(\boldsymbol{D}_L - \mathrm{diag}(\boldsymbol{\varepsilon}_r))\boldsymbol{E}_j(\boldsymbol{\varepsilon}_r) = \boldsymbol{b}_j. \tag{4}$$

Here, $\boldsymbol{D}_L$ is the finite difference matrix for the three-dimensional vector electrical field, representing the operator $\frac{1}{k_0}\nabla \times \nabla \times$ with perfectly matched layers (PMLs) on the boundary of the solution domain [25]. $\mathrm{diag}(\boldsymbol{\varepsilon}_r)$ represents a diagonal matrix constructed from the vectorized $\boldsymbol{\varepsilon}_r$. k_0 is the wavenumber in vacuum. $\boldsymbol{b}_j$ denotes the input excitation that induces the field $\boldsymbol{E}_j$ within the device and is derived from the fundamental TE mode of the input waveguide based on the total field/scattered field technique [26].

Combining Eqs. (3) and (4), the gradient of the objective function with respect to $\boldsymbol{\varepsilon}_r$ can be derived as

$$\nabla l(\boldsymbol{\varepsilon}_r) = \sum_{j=1}^{J} 2\,\mathrm{Re}\{\mathrm{diag}(\boldsymbol{E}_j^*)(\boldsymbol{D}_L - \mathrm{diag}(\boldsymbol{\varepsilon}_r))^{-1}\boldsymbol{E}_{tj}\}. \tag{5}$$

The inverse problems $(\boldsymbol{D}_L - \mathrm{diag}(\boldsymbol{\varepsilon}_r))^{-1}\boldsymbol{E}_{tj}$ and $(\boldsymbol{D}_L - \mathrm{diag}(\boldsymbol{\varepsilon}_r))^{-1}\boldsymbol{b}$ were both solved using the least squares method [27], and were carried out on $J = 4$ parallel GPUs (NVIDIA RTX 3090). The relative permittivity $\boldsymbol{\varepsilon}_r$ is updated along the gradient direction with an adaptive step size τ as $\boldsymbol{\varepsilon}_r \leftarrow \boldsymbol{\varepsilon}_r + \tau\nabla l(\boldsymbol{\varepsilon}_r)$. To promote the binary medium (air and silicon) on the silicon layer, the updated $\boldsymbol{\varepsilon}_r$ is mapped by a sigmoid function to produce the relative permittivity in the next iteration, and is expressed as

$$\boldsymbol{\varepsilon}_r' = \frac{\varepsilon_{\mathrm{Si}} - \varepsilon_{\mathrm{air}}}{1 + \exp\left(-\gamma\left(\boldsymbol{\varepsilon}_r - \frac{\varepsilon_{\mathrm{air}} + \varepsilon_{\mathrm{Si}}}{2}\right)\right)} + \varepsilon_{\mathrm{air}}. \tag{6}$$

Here, $\varepsilon_{\mathrm{Si}}$ and $\varepsilon_{\mathrm{air}}$ are the relative permittivity of silicon and air, respectively, and $\gamma = 4$ is a hyperparameter that controls the slope of the sigmoid function.

The inversely designed photonic dot-product core was fabricated on commercially available silicon-on-insulator (SOI) wafers. The wafers consisted of 250 nm silicon on top of a 3 μm buried oxide. The core was fabricated using a positive tone ZEP resist followed by electron beam lithography and inductively coupled plasma reactive ion etching. To realize the subwavelength sized and spaced features of the inversely designed structure, short range proximity correction was used to vary the dose of the exposure across the device. The core consisted of four single-mode input waveguides (480 nm in width) and two few-mode output waveguides (774 nm in width). The two few-mode output waveguides were each tapered to a 40 μm × 40 μm photonic crystal structure [28,29], which vertically couples out the electric field profiles for observation by a microscope imaging system. Details of the photonic crystal design and simulation results can be found in Appendix B.

A microscope setup was used to experimentally characterize the fabricated vector dot-product core. The core was

edge-coupled to the fiber array that provided four modulated inputs, each driven by an independent off-chip Mach–Zehnder modulator (MZM, JDSU IOAP-MOD9140). The modulators were driven by a multichannel digital-to-analog converter (DAC, Analog Devices, MAX11300), which was controlled by a microcontroller (Analog Devices, SDP-CK1Z). The modulated signals were edge-coupled into the four input ports of the dot-product core. The intensity profiles on the two vertical output couplers were recorded from above through a long working distance 20× objective and a tube lens onto a short-wave infrared (SWIR) camera (Allied Vision, Goldeye CL-008 TEC1). The camera and DAC synchronously perform 100 multiplications per second at the frame rate of the SWIR camera.

3. EXPERIMENTAL RESULTS

A. Characterization of the Fabricated Dot-Product Core

Figure 3(a) shows a microscope image of the photonic core under our characterization setup. Figure 3(b) plots the intensity profiles at the output coupler when the first two single-mode input arms, a_1 and a_2, were individually activated in the experiments. The intensity profiles match the target spatial profiles of the fundamental and second-order modes. To quantify the computing performance, we simulated the electromagnetic (EM) behavior of the designed and fabricated core using Ansys Lumerical FDTD software based on the design and scanning electron microscope (SEM) image, respectively. TE fundamental modes were launched into each single-mode input waveguide, and the resulting field profiles at a nominal operating wavelength of 1570 nm are shown in Figs. 3(c) and 3(d). The orange boxes show the cross-section of the electrical field profiles marked by the dashed lines.

The transfer matrices S_t of the designed and fabricated cores can be calculated from the overlap integral [30] between the cross-sectional electrical fields and the two TE eigenmodes in the top and bottom arms. Both matrices share the same structure as the ideal transfer matrix S_i in Eq. (2). The ideal inversely designed core features a symmetric design with <10% crosstalk, as indicated by the off-diagonal elements. The power-splitting ratios between the top and bottom arms are both approximately 46% versus 54% for the fundamental and second-order TE modes. The fabricated core maintains the relative low crosstalk with a maximum of 13.5% in the off-diagonal elements. The power splitting ratios are 41% versus 59% and 49% versus 51% for fundamental and second-order TE modes, respectively.

The insertion loss and crosstalk of the designed and fabricated core can be quantified by the crosstalk matrix M_X, whose elements are the overlap between the columns in the transfer matrices of the ideal (S_i) and the designed (or fabricated) device (S_t), expressed as

$$M_X[i,j] = S_t[:,i]^* \cdot S_i[:,j]. \tag{7}$$

Here, $[:, i]$ extracts the i-th column vector from the matrix. The insertion loss (IL) and crosstalk (XT) can both be derived from M_X, respectively, as [30]

$$\begin{aligned} \mathrm{IL(dB)} &= -10\log_{10}(\text{max eigenvalue of } M_X); \\ \mathrm{XT(dB)} &= -10\log_{10}\left(\frac{\text{power in the diagonals of } M_X}{\text{power in the off-diagonals of } M_X}\right). \end{aligned} \tag{8}$$

The crosstalk matrices at a normal operating wavelength of 1570 nm of both ideal and fabricated dot-product core designs are shown in Fig. 4(a). Figure 4(b) plots the insertion loss and crosstalk of the designed and fabricated cores as a function of the wavelength. The ideal dot-product core design features a consistent 2.3 dB insertion loss and a crosstalk of < −13 dB (<5%) across the wavelength range of 1540 nm to 1590 nm. The fabricated core maintains a consistent insertion loss and crosstalk within the wavelength range 1550 nm to 1580 nm, suggesting broadband performance that supports wavelength multiplexed inputs. Despite the uneven splitting of the input fields into the upper and lower arms, the crosstalk between the two spatial modes in the output waveguides is −9.06 dB, or 12.4%. The low crosstalk allows us to empirically correct most of the computing errors, as described in Appendix A.

B. General-Purpose Computing Examples

The core supports dot products between two-element vectors with fixed-point precision, enabling the deployment of general

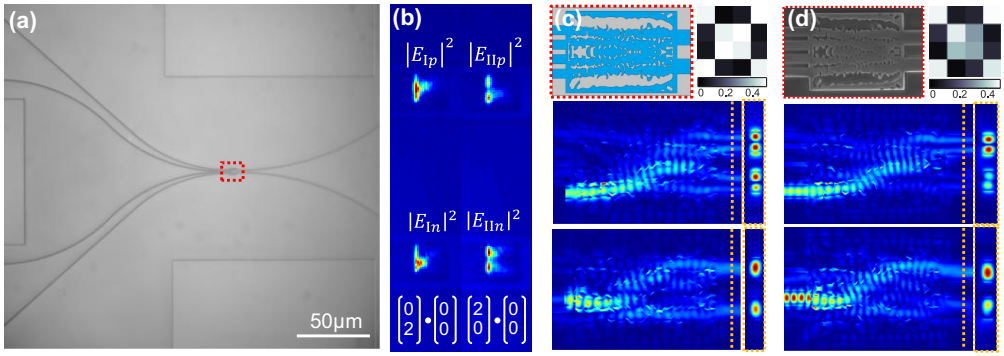

Fig. 3. Characterization of the fabricated dot-product core. (a) Microscope image of the fabricated dot-product core under test. (b) Experimentally observed intensity profiles on the two output couplers when the inputs a_1 and a_2 were individually excited. (c) Structure of the ideal inversely designed dot-product core and simulated electrical field profiles within the core. (d) SEM image of the fabricated dot-product core and simulated electrical field profiles within the core based on the SEM image. The side views show the electrical field profiles at the location marked by the orange dashed line.

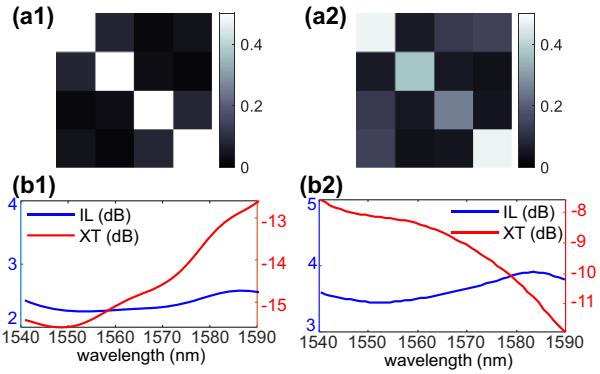

Fig. 4. Characterization of (a1), (b1) designed and (a2), (b2) fabricated dot-product core. (a) Crosstalk matrix M_X of the core. (b) Insertion loss and crosstalk (in dB) as a function of wavelength.

purpose computing tasks such as complex number multiplication and optical flow calculation. To carry out general purpose dot-product calculations, $(a_1, a_2) \cdot (b_1, b_2)^T$, on the photonic core, we calibrated the four MZMs to generate five signed linear analog levels representing the integers from −2 to 2 on each input. The four input ports of the inversely designed photonic structure receive the modulated optical signal representing a_1, a_2, b_2, and b_1, respectively, from top to bottom. The intensity differences between two output couplers were proportionally mapped to the dot products using the output from $(1, 0) \cdot (1, 0)^T$. Figure 5(a) plots a time-division multiplexing (TDM) sequence of 16 dot products performed on the photonic core. We quantify the computing error with normalized mean square error (NMSE) between the ground truth Y_{gt} and the experimental Y_{exp} dot products, defined in

$$\mathrm{NMSE} = \frac{\sum_{k=1}^{K} |Y_{k,\mathrm{exp}} - Y_{k,\mathrm{gt}}|^2}{\sum_{k=1}^{K} |Y_{k,\mathrm{gt}}|^2}. \tag{9}$$

Here, the summation is performed over all K symbols in the sequence. The NMSE of all multiplications was 6.32%, offering sufficient dynamic range to represent signed integers from −8 to 8 (signed 4-bits) in the dot-product results.

We first applied the photonic dot-product core to perform complex number multiplication [i.e., $(a + bi) \times (c + di)$]. The real and imaginary parts of the result $[(ac - bd) + (ad + bc)i]$ are split into two equivalent dot products encoded in a TDM symbol sequence. Sixteen complex number pairs represented by a sequence of 32 dot products were multiplied on the core. Figure 5(c) compares the products from the ideal and fabricated cores with the ground truth on the complex plane. The designed dot-product core shows good agreement with ground truth and an NMSE of 4.0%, suggesting that the design can reach a dynamic range of signed 25 levels, or greater than the signed 4-bit precision. The NMSE between the ground truth and experimental complex products is 15.9%, which is consistent with the simulation of a fabricated dot-product core. The computing error is primarily attributed to the fabrication deviation from the ideal design and the time-varying phase instability from the off-chip fiber inputs. The phase stability can be improved by switching to on-chip modulators. The fabrication deviation can be compensated with additional phase modulation on each input, which can be generated from integrated thermal optical phase shifters.

In addition, we have also demonstrated a computer vision task using the photonic dot-product core. Specifically, we use the device to calculate the optical flow in a visual scene to quantify the motion of the object. The real-time calculation of the

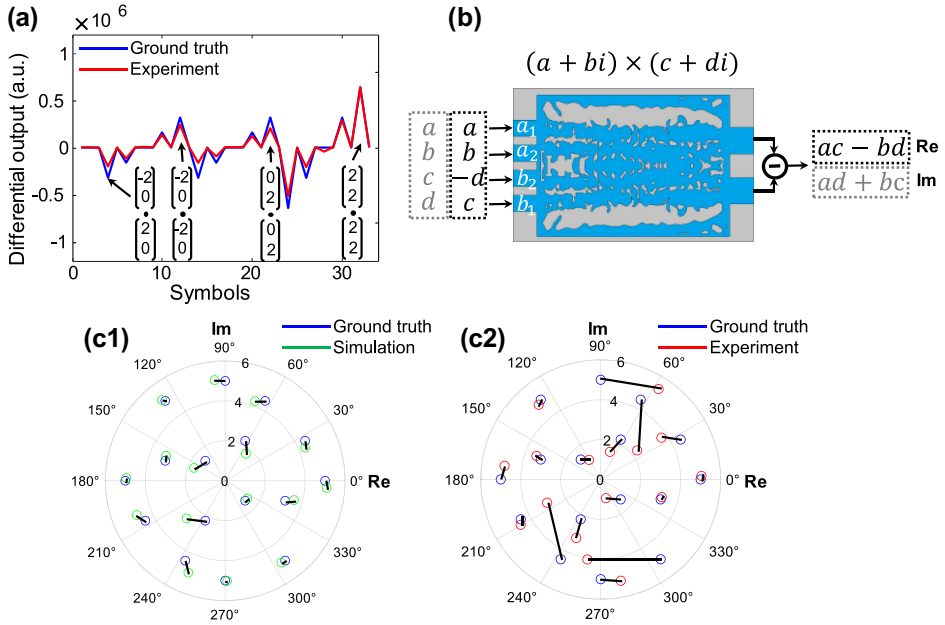

Fig. 5. General-purpose computing examples as dot-products on the photonic core. (a) Dot-product calculation of a sequence of 16 two-element vectors. (b) Complex number multiplications encoded as two equivalent dot-products in time-division multiplexing. (c1), (c2) Multiplication results between 16 complex numbers. Blue circles indicate ground truth results, green circles indicate simulated results from the ideal inversely designed core in (b), and red circles indicate experimental results calculated on the fabricated dot-product core.

optical flow in a dynamic environment plays an important role in motion detection and object tracking of computer vision systems [31,32]. Here, we calculated the optical flow of selected edge pixels between two adjacent two-dimensional frames, $I_1(x,y)$ and $I_2(x,y)$, from a 10 frames-per-second spinning wheel animation on the dot-product core. The flow vector $(u,v)^T$ satisfies $(d_x,d_y)\cdot(u,v)^T=-d_t$, where d_x, d_y, and d_t are the finite differences of the image $I_t(x,y)$ along x, y, and t dimensions, respectively [33]. Due to the ambiguity in uniquely determining the pixelwise $(u,v)^T$, we expand the optical flow vector onto the diagonal pixels in a 2×2 window, as

$$\begin{bmatrix} d_{x11} & d_{y11} \\ d_{x22} & d_{y22} \end{bmatrix}\cdot\begin{bmatrix} u \\ v \end{bmatrix}=-\begin{bmatrix} d_{t11} \\ d_{t22} \end{bmatrix}. \tag{10}$$

Assuming uniform flow vectors in the 2×2 window, the calculations are broken down into two parts: (i) on the two pixels [marked in gray in Fig. 6(b)] along the primary diagonals d_{x11}, d_{x22}, d_{y11}, d_{y22}, d_{t11}, and d_{t22}; and (ii) along the secondary diagonals d_{x12}, d_{x21}, d_{y12}, d_{y21}, d_{t12}, and d_{t21} [marked in white in Fig. 6(b)]. Results from the primary and secondary diagonals are averaged to obtain the flow vector within the 2×2 window. Equation (10) can be solved using Cramer's rule, written as

$$u=-\frac{\begin{vmatrix} d_{t11} & d_{y11} \\ d_{t22} & d_{y22} \end{vmatrix}}{\begin{vmatrix} d_{x11} & d_{y11} \\ d_{x22} & d_{y22} \end{vmatrix}},\quad v=-\frac{\begin{vmatrix} d_{x11} & d_{t11} \\ d_{x22} & d_{t22} \end{vmatrix}}{\begin{vmatrix} d_{x11} & d_{y11} \\ d_{x22} & d_{y22} \end{vmatrix}}. \tag{11}$$

Here, all the 2×2 determinants $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ are computed as their equivalent dot products $ad-bc$. The flow vector within one 2×2 window requires six dot-product calculations encoded in a TDM sequence.

Figure 6(c) shows the optical flow vector within eight ($J=8$) 2×2 windows on the edges of the spinning wheel. We quantify the error in flow vector calculation using the mean cosine similarity, defined in

$$S_c=\frac{1}{J}\sum_{j=1}^{J}\frac{(u_{j,\text{ideal}},v_{j,\text{ideal}})\cdot(u_{j,\text{exp}},v_{j,\text{exp}})}{\|(u_{j,\text{ideal}},v_{j,\text{ideal}})\|\|(u_{j,\text{exp}},v_{j,\text{exp}})\|}. \tag{12}$$

Here, $(u_{j,\text{ideal}},v_{j,\text{ideal}})$ and $(u_{j,\text{exp}},v_{j,\text{exp}})$ represent the ideal flow vector and the one calculated on the fabricated dot-product core, respectively. The mean cosine similarity is 81.8%, suggesting that the flow vectors calculated on the photonic dot-product core have captured the correct rotation direction. The mean magnitude of the flow vectors reflects the angular speed of the wheel, which is 1.32 rad/s based on the calculated optical flow on the fabricated dot-product core. Compared to the ground truth 1.41 rad/s, the relative error of the angular speed calculation is 6.8%. This example illustrates that the fixed-point dot-product core can be used in conjunction with a tailored algorithm to extract features-of-interest in computer vision tasks.

4. CONCLUSION

In summary, we have demonstrated a compact, integrated photonic dot-product core from an inverse design. The core utilizes spatial mode as the multiplexing dimension to perform arbitrary two-element vector dot products. To account for the difference between the design and fabricated photonic structures, calibration and error correction routines have been developed and tested on the fabricated dot-product core. We have demonstrated an equivalent signed 4-bit precision in the dot-product results, and successfully deployed a general purpose complex number multiplier and an optical flow calculator on the fabricated device.

The miniaturized footprint enables the large-scale integration of the core as part of the photonic primitives in an electronic-photonic co-packaged parallel computing array on modern CMOS-compatible platforms. Combining our current design with on-chip modulators and multimode photodiodes [34], a computing speed on the order of 10^9 dot products per second is supported by modern gigabaud optoelectronics. By further integrating dense wavelength-division multiplexing (DWDM) channels and spatial modes as super-dimensions in photonic matrix- and tensor-based processors [10], our strategy enables a computing throughput on the order of

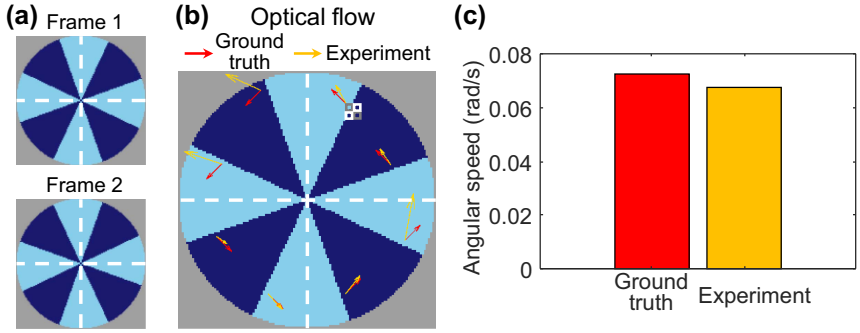

Fig. 6. Optical flow calculation between two adjacent frames of a spinning wheel animation. (a) Two frames from the spinning wheel animation with 100 ms interval (10 frames per second). (b) Optical flow vector of eight edge pixels. Red arrows indicate the ground truth of the flow vectors, and orange arrows indicate experimental results calculated on the photonic dot-product core. (c) Comparison of the calculated angular speed on the dot-product core with ground truth.

10^3 TOPS/mm^2, which is orders of magnitude higher than that of dedicated electronic vector/matrix accelerators [4,35].

APPENDIX A: CALIBRATION AND COMPUTING ERROR OF DOT-PRODUCT CORE

Given the transfer matrix S_t of the fabricated dot-product core, it is possible to compensate the four inputs to account for the uneven splitting and/or crosstalk due to fabrication imperfections. This process mixes the four inputs according to the compensation matrix C before sending to the dot-product core. The choice of C must minimize the crosstalk and equalize the amplitudes in the two output arms for both spatial modes. Theoretically, C can be calculated according to

$$S_t \cdot C = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 0 & i \\ 0 & 1 & i & 0 \\ 0 & i & 1 & 0 \\ i & 0 & 0 & 1 \end{bmatrix}. \tag{A1}$$

Here, the right-hand side denotes the transfer matrix of an ideal dot-product core in Eq. (2). Figure 7 shows the transfer matrix of the fabricated core and the corresponding compensation matrix $\tilde{C}_{\text{full}}$ for pre-mixing the four inputs. However, using the full transfer matrix for compensation involves the multiplication of a 4×4 complex matrix on top of the desired inputs, giving rise to 16 additional digital MAC operations.

Because of the low crosstalk among the four output field profiles, we can also empirically treat two spatial modes independently and ignore the off-diagonal elements in the transfer matrix when correcting the dot-products in experiments. This is equivalent to constraining all the off-diagonal elements in S_t and its associated compensation matrix $\tilde{C}_{\text{emp}}$ to zero. The approximated transfer matrix of the fabricated core is shown in

$$\begin{bmatrix} E_{\text{II}p} \\ E_{\text{I}p} \\ E_{\text{I}n} \\ E_{\text{II}n} \end{bmatrix} = \begin{bmatrix} S_{\text{II}11} & 0 & 0 & S_{\text{II}12} \\ 0 & S_{\text{I}11} & S_{\text{I}12} & 0 \\ 0 & S_{\text{I}21} & S_{\text{I}22} & 0 \\ S_{\text{II}21} & 0 & 0 & S_{\text{II}22} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ b_2 \\ b_1 \end{bmatrix}. \tag{A2}$$

Based on Eq. (A2), the empirical compensation can be formulated as a complex matrix [Eq. (A1)] in which only two elements in each row are nonzero. It is worth noting that experimentally implementing a complex compensation matrix $\tilde{C}_{\text{emp}}$ entails generating accurate phase modulations on individual inputs, which are hindered by the intrinsic time-varying phase on the four off-chip inputs.

To perform empirical compensation based solely on amplitude scaling, we observe the intensity difference between the two output couplers from Eq. (A2), expressed as

$$\begin{aligned} I_{\text{diff}} = {} & |a_2|^2(|S_{\text{I}11}|^2 - |S_{\text{I}21}|^2) + |b_2|^2(|S_{\text{I}12}|^2 - |S_{\text{I}22}|^2) \\ & + 2\text{Re}\{a_2 b_2^*(S_{\text{I}11}S_{\text{I}12}^* - S_{\text{I}21}S_{\text{I}22}^*)\} \\ & + |a_1|^2(|S_{\text{II}11}|^2 - |S_{\text{II}21}|^2) + |b_1|^2(|S_{\text{II}12}|^2 - |S_{\text{II}22}|^2) \\ & + 2\text{Re}\{a_1 b_1^*(S_{\text{II}11}S_{\text{II}12}^* - S_{\text{II}21}S_{\text{II}22}^*)\}. \end{aligned} \tag{A3}$$

Equation (A3) indicates that to produce the result proportional to the dot product $a_1b_1 + a_2b_2$, the residue terms $|a_2|^2(|S_{\text{I}11}|^2 - |S_{\text{I}21}|^2) + |b_2|^2(|S_{\text{I}12}|^2 - |S_{\text{I}22}|^2) + |a_1|^2(|S_{\text{II}11}|^2 - |S_{\text{II}21}|^2) + |b_1|^2(|S_{\text{II}12}|^2 - |S_{\text{II}22}|^2)$ arising from the uneven splitting can be corrected by removing the differential outputs when only one of the four inputs is excited. The amplitude scaling factors $\text{Re}\{S_{\text{I}11}S_{\text{I}12}^* - S_{\text{I}21}S_{\text{I}22}^*\}$ and $\text{Re}\{S_{\text{II}11}S_{\text{II}12}^* - S_{\text{II}21}S_{\text{II}22}^*\}$ can be merged into the calibrated amplitude of b_1 and b_2.

Based on the transfer matrix S_t of the fabricated core, Fig. 8 simulates the computing errors before and after compensation with 10 40-symbol random sequences of dot products deployed on the fabricated dot-product core. Both compensation methods are effective in reducing computing errors by at least two-fold, giving a nominal NMSE of ~15% after empirical compensation. The empirical compensation with amplitude scaling only involves remapping the amplitude of the individual

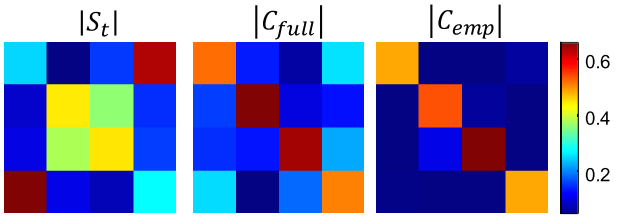

Fig. 7. Transfer matrix of the fabricated dot-product core S_t and the corresponding compensation matrices C_{full} and C_{emp}. Only the magnitudes of the matrix elements are shown.

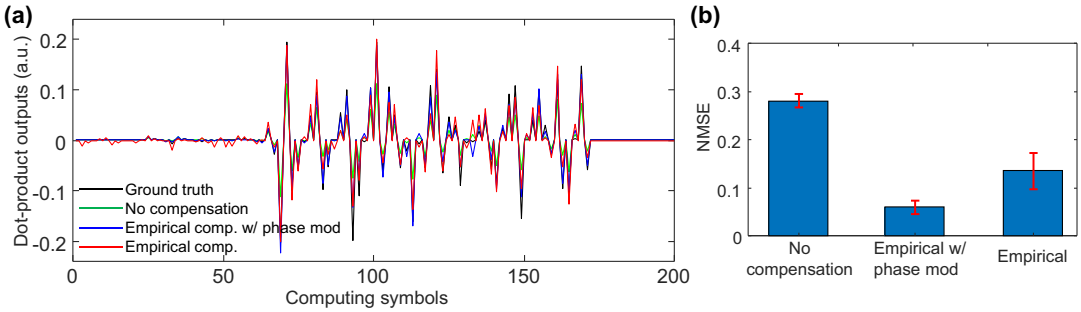

Fig. 8. Comparison of the different compensation methods. (a) TDM dot-product output sequence before and after compensation using the full transfer matrix S_t and the empirical method with/without phase modulation. (b) Comparison between the NMSE of the raw and compensated dot products.

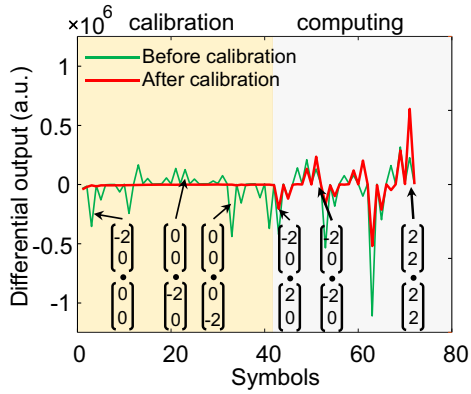

Fig. 9. Comparison of the dot products before and after calibration in experiments.

inputs, which can be efficiently implemented as an input lookup table on a microcontroller and is thus computationally efficient. In the experiments, we appended all the single-input excitations as a calibration header and used them to remove residue terms from the computing symbols, as shown in Fig. 9.

The NMSE after the empirical compensation presented here represents a theoretical lower bound in the computing error. In actual experiments, the time-varying phase on the four off-chip modulated inputs cannot be measured and compensated. As a result, the experimental computing error could be higher than the lower bound. We envision that with fully integrated optical paths on a chip, including the use of on-chip modulators [36] and few-mode photodiodes [34], the time-varying phase could be resolved.

APPENDIX B: CHARACTERIZATION OF HIGHER-ORDER SPATIAL MODES

The intensity profiles of different spatial modes on the output multimode waveguide can be coupled vertically using a photonic crystal structure acting as a high-order grating coupler. In our design, the photonic crystal structure measures 40 μm × 40 μm with a pitch of 0.64 μm and a hole size of 0.32 μm. The two multimode output waveguides from our inversely designed structure are each tapered to a photonic crystal for observing the intensity profiles by a microscope imaging system from above. Figure 10 shows the intensity profiles of different input modes and the output on top of the photonic crystal region from FDTD simulation. The number of lobes in the intensity profile indicates the order of the TE modes.

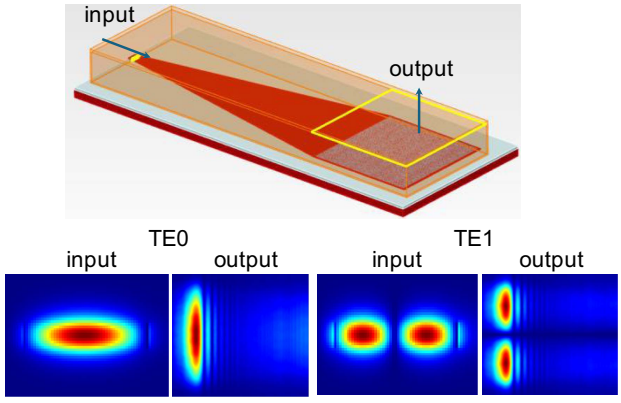

Fig. 10. Design and FDTD simulations of the photonic crystal output coupler supporting the observation of multimode intensity profiles from the top.

Funding. Office of Naval Research (N00014-20-1-2441); Army Research Office (W911NF2110321); National Science Foundation (1932858).

Acknowledgment. R.S. acknowledges the support of the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration. This work was performed in part at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are available in the text. Raw data may be obtained from the authors upon reasonable request.

REFERENCES

1. L. S. Blackford, A. Petitet, R. Pozo, *et al.*, "An updated set of basic linear algebra subprograms (BLAS)," ACM Trans. Math. Softw. **28**, 135–151 (2002).
2. M. J. Flynn, "Some computer organizations and their effectiveness," IEEE Trans. Comput. **100**, 948–960 (1972).
3. H. Kaul, M. A. Anders, S. K. Mathew, *et al.*, "A 300 mV 494GOPS/W reconfigurable dual-supply 4-way SIMD vector processing accelerator in 45 nm CMOS," IEEE J. Solid-State Circuits **45**, 95–102 (2010).
4. S. K. Hsu, A. Agarwal, M. A. Anders, *et al.*, "A 280 mV-to-1.1 V 256b reconfigurable SIMD vector permutation engine with 2-dimensional shuffle in 22 nm tri-gate CMOS," IEEE J. Solid-State Circuits **48**, 118–127 (2013).
5. Y. Shen, N. C. Harris, S. Skirlo, *et al.*, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441–446 (2017).
6. R. Hamerly, L. Bernstein, A. Sludds, *et al.*, "Large-scale optical neural networks based on photoelectric multiplication," Phys. Rev. X **9**, 021032 (2019).
7. B. J. Shastri, A. N. Tait, T. F. de Lima, *et al.*, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**, 102–114 (2021).
8. K. Kikuchi, "Fundamentals of coherent optical fiber communications," J. Lightwave Technol. **34**, 157–179 (2016).
9. B. Murmann, "ADC performance survey 1997–2022," (2022), http://www.stanford.edu/~murmann/adcsurvey.html.
10. A. Fardoost, F. G. Vanani, Z. Zhu, *et al.*, "A high-speed photonic tensor accelerator," in *IEEE Photonics Conference (IPC)* (IEEE, 2022), pp. 1–2.
11. Z. Zhu, A. Fardoost, F. G. Vanani, *et al.*, "Coherent general-purpose photonic matrix processor," ACS Photon. **11**, 1189–1196 (2024).

12. M. A. Nahmias, T. F. de Lima, A. N. Tait, *et al.*, "Photonic multiply-accumulate operations for neural networks," IEEE J. Sel. Top. Quantum Electron. **26**, 7701518 (2020).
13. A. N. Tait, M. A. Nahmias, B. J. Shastri, *et al.*, "Broadcast and weight: an integrated network for scalable photonic spike processing," J. Lightwave Technol. **32**, 3247–3439 (2014).
14. T. F. de Lima, H.-T. Peng, A. N. Tait, *et al.*, "Machine learning with neuromorphic photonics," J. Lightwave Technol. **37**, 1515–1534 (2019).
15. X. Wu, C. Huang, K. Xu, *et al.*, "Mode-division multiplexing for silicon photonic network-on-chip," J. Lightwave Technol. **35**, 3223–3228 (2017).
16. K. Y. Yang, C. Shirpurkar, A. D. White, *et al.*, "Multi-dimensional data transmission using inverse-designed silicon photonics and microcombs," Nat. Commun. **13**, 7862 (2022).
17. D. Dai, J. Wang, and Y. Shi, "Silicon mode (de)multiplexer enabling high capacity photonic networks-on-chip with a single-wavelength-carrier light," Opt. Lett. **38**, 1422–1424 (2013).
18. G. Zhang, H. R. Mojaver, A. Das, *et al.*, "Mode insensitive switch for on-chip interconnect mode division multiplexing systems," Opt. Lett. **45**, 811–814 (2020).
19. H. Shiran, G. Zhang, and O. Liboiron-Ladouceur, "Dual-mode broadband compact 2×2 optical power splitter using sub-wavelength metamaterial structures," Opt. Express **29**, 23864–23876 (2021).
20. Y. Zhang, M. A. Al-Mumin, H. Liu, *et al.*, "An integrated few-mode power splitter based on multimode interference," J. Lightwave Technol. **37**, 3000–3008 (2019).
21. L. Su, D. Vercruysse, J. Skarda, *et al.*, "Nanophotonic inverse design with SPINS: software architecture and practical considerations," Appl. Phys. Rev. **7**, 011407 (2020).
22. C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, *et al.*, "Adjoint shape optimization applied to electromagnetic design," Opt. Express **21**, 21693–21701 (2013).
23. D. Melati, Y. Grinberg, M. Kamandar Dezfouli, *et al.*, "Mapping the global design space of nanophotonic components using machine learning pattern recognition," Nat. Commun. **10**, 4775 (2019).
24. S. Molesky, Z. Lin, A. Y. Piggott, *et al.*, "Inverse design in nanophotonics," Nat. Photonics **12**, 659–670 (2018).
25. W. Shin and S. Fan, "Choice of the perfectly matched layer boundary condition for frequency-domain Maxwell's equations solvers," J. Comput. Phys. **231**, 3406–3431 (2012).
26. R. C. Rumpf, *Electromagnetic and Photonic Simulation for the Beginner: Finite-Difference Frequency-Domain in MATLAB* (Artech House, 2022).
27. R. Barrett, M. Berry, T. F. Chan, *et al.*, "Templates for the solution of linear systems: building blocks for iterative methods," Math. Comput. **64**, 1349–1352 (1995).
28. Y. Tong, W. Zhou, X. Wu, *et al.*, "Efficient mode multiplexer for few-mode fibers using integrated silicon-on-insulator waveguide grating coupler," IEEE J. Quantum Electron. **56**, 8400107 (2020).
29. B. Wohlfeil, G. Rademacher, C. Stamatiadis, *et al.*, "A two-dimensional fiber grating coupler on SOI for mode division multiplexing," IEEE Photon. Technol. Lett. **28**, 1241–1244 (2016).
30. N. K. Fontaine, R. Ryf, H. Chen, *et al.*, "Design of high order mode-multiplexers using multiplane light conversion," in *European Conference on Optical Communication (ECOC)* (IEEE, 2017), pp. 1–3.
31. Y. Mae, Y. Shirai, J. Miura, *et al.*, "Object tracking in cluttered background based on optical flow and edges," in *13th International Conference on Pattern Recognition* (IEEE, 1996), Vol. **1**, pp. 196–200.
32. Z. Chen, J. Cao, Y. Tang, *et al.*, "Tracking of moving object based on optical flow detection," in *International Conference on Computer Science and Network Technology* (IEEE, 2011), Vol. **2**, pp. 1096–1099.
33. B. K. P. Horn and B. G. Schunck, "Determining optical flow," Artif. Intell. **17**, 185–203 (1981).
34. L. Wu, D. Lv, N. Zhao, *et al.*, "Research on germanium photodetector with multi-mode waveguide input," Photonics **10**, 455 (2023).
35. A. Amirsoleimani, F. Alibart, V. Yon, *et al.*, "In-memory vector-matrix multiplication in monolithic complementary metal–oxide–semiconductor-memristor integrated circuits: design choices, challenges, and perspectives," Adv. Intell. Syst. **2**, 2000115 (2020).
36. J. Sun, R. Kumar, M. Sakib, *et al.*, "A 128 Gb/s PAM4 silicon microring modulator with integrated thermo-optic resonance tuning," J. Lightwave Technol. **37**, 110–115 (2019).