





# Highly efficient photonic convolver via lossless mode-division fan-in

Received: 17 April 2025

Accepted: 6 August 2025

Published online: 13 August 2025

Shangsen Sun<sup>1,3</sup>, Shiji Zhang<sup>1,3</sup> , Bo Wu<sup>1</sup>, Shan Jiang<sup>1</sup>, Baiheng Zhao<sup>1</sup>,  
Hailong Zhou<sup>1</sup> , Jianji Dong<sup>1,2</sup> ✉ & Xinliang Zhang<sup>1,2</sup> 

Optical neural networks (ONNs) leverage the parallelism and low-energy consumption of photonic signal processing to overcome the limitations of traditional electronic computing. Optics inherently enables fan-in and fan-out without the Resistor-Capacitor (RC) and Inductor-Capacitor (LC) delays of electrical interconnects. However, for single-mode photonic integrated circuits, reciprocity constraints introduce unavoidable loss during beam combining, hindering large-scale on-chip photonic fan-in. To overcome this challenge, we provide a photonic lossless mode-division fan-in solution for the convolution accelerators. Using inverse design, we developed a compact multimode photonic convolution accelerator ( $0.42\text{ mm}^2$ ) with  $\pm 15\text{ nm}$  fabrication tolerance and  $35\text{ nm}$  optical bandwidth, enabling parallel computation across mode and wavelength dimensions. Experimental results in the C-band confirm a 6–7 bit convolution precision, leading to classification accuracies of 95.2% on MNIST and 87.9% on Fashion-MNIST. Moreover, the device offers a theoretical computational density of  $125.14\text{ TOPS/mm}^2$ , underscoring its potential for scalable and energy-efficient photonic computing accelerators.

Deep neural networks (DNNs) have driven remarkable advances in various artificial intelligence (AI) applications<sup>1</sup>, yet their ever-growing computational demands strain conventional electronic hardware<sup>2,3</sup>. To sustain performance scaling, new hardware paradigms are needed to overcome interconnect, power, and parallelism bottlenecks. Optics offers a promising route to overcome these constraints by leveraging its intrinsic parallelism for signal processing<sup>4–8</sup>.

Specifically, summing outputs from multiple parallel-processing units (fan-in) is a core operation in computing architectures<sup>9</sup>, traditionally requiring multiple buffering stages in electronics to address Resistor-Capacitor (RC) and Inductor-Capacitor (LC) delays<sup>10,11</sup>. By contrast, optical systems can merge multiple beams in free space (using lenses<sup>12</sup>, holographic elements<sup>13</sup>, or intensifiers<sup>14</sup>) or on chip (using multimode interferometers<sup>15</sup> or directional couplers<sup>16</sup>) without incurring the same interconnect overheads. However, for single-mode photonic integrated circuits, reciprocity imposes fundamental insertion losses during beam combining: merging multiple inputs into one single-mode waveguide typically causes partial back-

reflection or scattering<sup>17</sup>, limiting the scalability of photonic computing systems. In particular, convolution operations in DNNs place a heavy load on such fan-in processes<sup>18–24</sup>, as they involve repeated additions of partial products. This makes convolution especially reliant on efficient fan-in, underscoring the importance of a low-loss photonic approach.

Here, we propose a photonic convolution accelerator featuring a lossless mode-division fan-in architecture that addresses the fundamental limitations of single-mode photonic computing. By exploiting the orthogonality of distinct waveguide modes, multiple signals are merged without incurring the insertion losses characteristic of conventional single-mode fan-in (SFI). Through inverse design, we have developed a high-tolerance ( $\pm 15\text{ nm}$ ) multimode 3-dB coupler and a low-power ( $23.6\text{ mW}$ ) multimode kernel modulator (MKM) with a 12 dB extinction ratio in the C-band, both supporting three modes and readily scalable to higher mode counts. To validate our approach, we focus on the convolution operation as a demonstration of our design. Leveraging parallel computation across both mode

<sup>1</sup>Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China. <sup>2</sup>Optics Valley Laboratory, 430074 Wuhan, China. <sup>3</sup>These authors contributed equally: Shangsen Sun, Shiji Zhang. ✉e-mail: [jjdong@mail.hust.edu.cn](mailto:jjdong@mail.hust.edu.cn)

and wavelength dimensions, our system achieves 6-7 bits of computational accuracy in the C-band, yielding classification accuracies of 95.2% on MNIST and 87.9% on Fashion MNIST. Moreover, we attain a theoretical arithmetic density of 125.14 TOPS/mm<sup>2</sup>, underscoring the potential of this solution to pave the way for scalable, energy-efficient photonic computing.

## Results

### Principle

Figure 1a illustrates the conceptual difference between traditional SFI and our multimode fan-in (MFI), showing their respective scattering matrices for a 3-port device with two inputs and one output. In single-mode photonic integrated circuits, signals from multiple waveguides converge into a single waveguide mode. Due to the reciprocity of passive photonic devices, the scattering matrix  $S$  remains symmetric, i.e.,  $S_{\mu\nu} = S_{\nu\mu}$ . Consequently, merging multiple inputs into a single output inevitably induces partial back-reflections or out-of-phase interference, causing intrinsic insertion losses. In practice, this diminishes the signal-to-noise ratio and restricts the scalability of on-chip fan-in. As an illustrative example, Fig. 1a (left) shows two separate inputs ( $a_1, a_2$ ) combining into a single output port ( $b_3$ ). The scattering

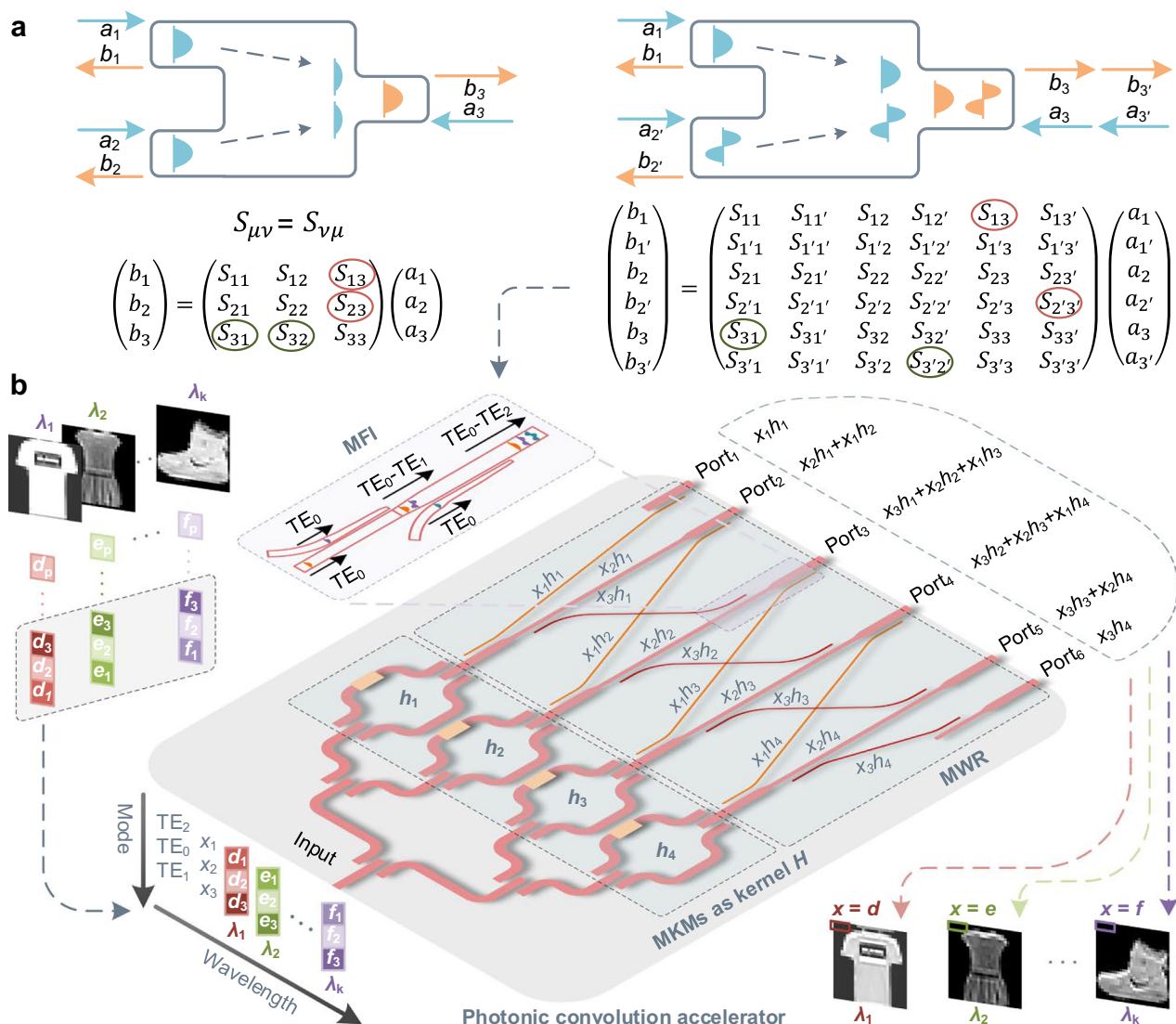
matrix is given by:

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \text{ with } S_{\mu\nu} = S_{\nu\mu} \quad (1)$$

Under the reciprocity constraints and  $|S_{13}|^2 + |S_{23}|^2 \leq 1$  (energy conservation), the output follows:

$$\frac{|S_{31}|^2 + |S_{32}|^2}{1^2 + 1^2} \leq \frac{1}{2} \quad (2)$$

indicating a minimum 50% fan-in loss for two input ports. As the number of inputs grows, these losses compound, ultimately constraining the scalability of on-chip fan-in. For  $N$  inputs, the output can contain at most  $1/N$  of the total input power. This limitation is fundamental to the incoherent fan-in of optical power, a scenario highly relevant to on-chip convolution accelerators. In such systems, achieving the convolution often involves parallel multiplications and routing signals through varied optical paths, which makes coherent



**Fig. 1 | Principle of a photonic multimode convolution accelerator enabled by lossless mode-division fan-in. a** The illustration of single-mode fan-in (SFI) and multimode fan-in (MFI) and their corresponding scattering matrix. **b** Design of the

photonic multimode convolution accelerator. MKM, multimode kernel modulator; MWR, multimode waveguide routing.

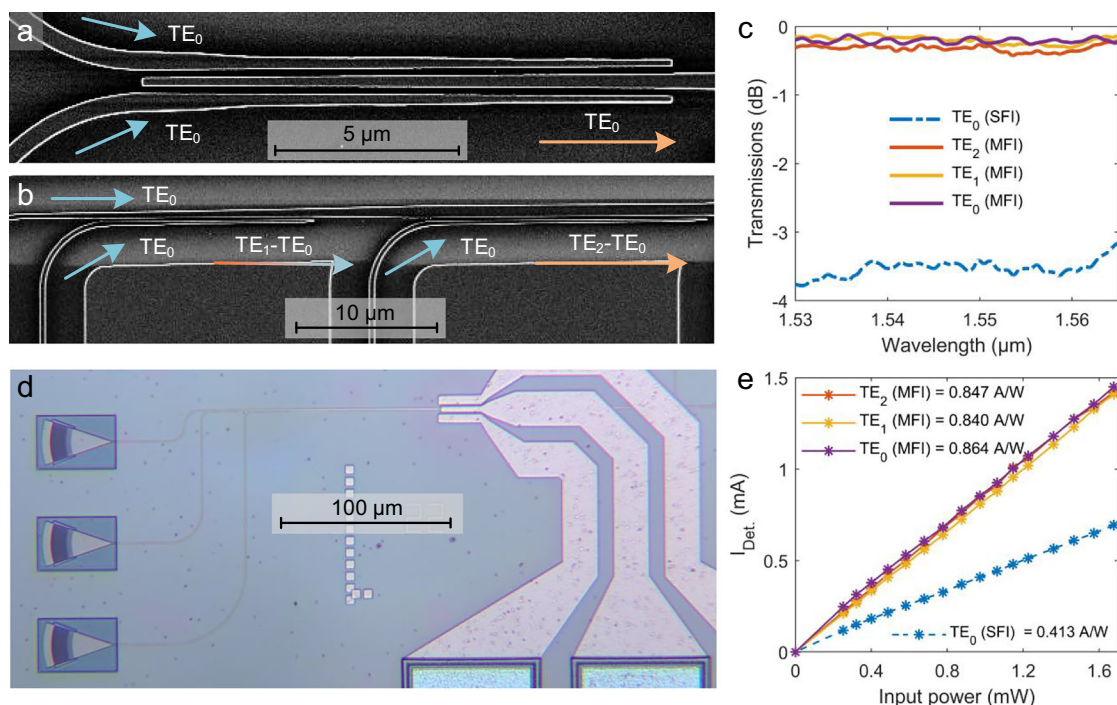
phase synchronization impractical, necessitating a direct power-summation approach.

By contrast, MFI makes use of multiple orthogonal waveguide modes to combine signals without forcing them to share the same mode profile. As shown in Fig. 1a, each input beam is mapped to a distinct spatial mode in the output waveguide, thus reducing interference and reflection between channels. The scattering matrix in the multimode design is still subject to reciprocity; however, distributing signals across orthogonal modes alleviates the inherent single-mode bottleneck. When carefully engineered, mode overlap can be minimized, allowing each mode to act as an independent channel for fan-in with negligible loss. This principle underpins our lossless mode-division fan-in architecture, where signals are effectively combined in a way that conserves total optical power and maximizes device utilization. Figure 1b illustrates our design of a multimode convolution accelerator enabled by lossless mode-division fan-in. The accelerator comprises multimode 3-dB couplers, MKMs, and multimode waveguide routing (MWR). The multimode waveguide is widened for clarity in the figure, and the structures of the multimode 3-dB couplers and MKMs are simplified for clear understanding (detailed structures are provided in Supplementary Notes 1 and 4). To process a single image, the image is first flattened into a one-dimensional vector of length  $P$ . The accelerator then receives  $M = 3$  vectors corresponding to  $M$  mode intensities, which propagate through cascaded multimode 3-dB couplers that split the energy into  $N = 4$  equal parts. Subsequently, the energy from each port propagates through  $N$  MKMs, which are modulated in intensity to function as a convolutional kernel  $H$ . The kernel remains fixed during each inference, rarely requiring rapid updates, so its response time does not impact the overall inference speed. The sliding output of the convolutional kernel window is then generated by MWR (the sliding path is represented by different colors), which comprises a mode demultiplexer and an MFI, with  $M + N - 1$  output ports. The convolution sliding process is entirely passive, with power consumption arising solely from input and kernel modulation. Additionally, each mode can transmit information across multiple wavelengths simultaneously, with the number of wavelengths ( $K$ )

corresponding to the number of images that can be input at once. Only  $M$  modulators are required to encode input vectors of size  $M$  and  $N$  MKMs to encode convolution kernels of size  $N$ , completing the computation in one clock cycle. The total number of multiply-accumulate (MAC) operations is  $K \times M \times N$ , which corresponds to  $2 \times K \times M \times N$  arithmetic operations.

### Multimode fan-in

MFI maps each input onto a distinct orthogonal waveguide mode, ensuring that signals do not interfere with one another. For the MFI to support multiple wavelengths simultaneously, it must exhibit broadband characteristics and high fabrication tolerance<sup>25–27</sup>. Inverse design<sup>28–30</sup>, which ensures both compactness and optimal performance, has emerged as a powerful tool in the development of various photonic components<sup>31–35</sup>. Figure 2 presents experimental results for high-performance SFI and MFI realized through inverse design. Figure 2a, b show scanning electron microscope (SEM) images of the SFI, with a length of 14.1  $\mu\text{m}$ , and the MFI for the  $\text{TE}_2$  and  $\text{TE}_1$  modes, with lengths of 21 and 20  $\mu\text{m}$ , respectively. Detailed parameters and design processes are provided in Supplementary Notes 1 and 2. Figure 2c shows the measured insertion losses over a broadband wavelength range. In the C-band, the SFI incurs an average loss of around 3.53 dB (dashed line), while the MFI exhibits average losses of only 0.32, 0.2, and 0.21 dB for  $\text{TE}_2$ - $\text{TE}_0$  modes (solid lines). To further verify the MFI results, we fabricated a Negative-Intrinsic-Positive (NIP) germanium-on-silicon vertical multimode photodetector (MPD) for three-mode detection. The MPD enables simultaneous reception and summation of multiple modes, whereas single-mode photodetectors require separate detection and summation. As shown in Fig. 2d, we integrated the MFI with the MPD, where the multimode waveguide is widened to 2  $\mu\text{m}$ , the covering Ge is widened to 4  $\mu\text{m}$ , and its length is set to 20  $\mu\text{m}$  to improve detector responsivity for all modes. The NIP-MPD has two metal contacts connected to the P-doped section to reduce attenuation near the peak of the modal power profiles. Figure 2e demonstrates the responsivities for the SFI and MFI. At 1.55  $\mu\text{m}$ , the measured responsivity is 0.413 A/W for the SFI and 0.847, 0.84, and 0.864 A/W



**Fig. 2 | Comparison of the SFI versus the MFI. a, b** Scanning electron microscope (SEM) images of the SFI and the MFI structures. **c** Measured insertion losses of SFI and MFI. **d** Optical microscope image of MFI cascaded with a multimode photodetector. **e** Responsivities for the SFI and MFI.



for the TE<sub>2</sub>-TE<sub>0</sub> modes of the MFI, respectively. In conclusion, the loss of the designed MFI is negligible compared to that of the SFI, demonstrating that distributing signals across orthogonal modes avoids the aggregation losses inherent to the SFI.

### Multimode kernel modulator

The input of the multimode convolution accelerator is split by a cascaded multimode 3-dB coupler, but conventional designs suffer from large footprints, limited mode scalability, and fabrication challenges<sup>36–39</sup>. To address these issues, we employed inverse design—specifically the Particle Swarm Optimization combined with Direct Range Search (PSO-DRS) optimization algorithm<sup>40</sup>—to develop a compact multimode 3-dB coupler composed of a multimode power splitter and the MFI (Supplementary Notes 1 and 2). Simulation results of the multimode power splitter indicate that with a  $\pm 15$  nm waveguide width variation, the insertion losses remain below 0.3 dB and the crosstalk stays below -20 dB. All devices, including subsequent components, were fabricated on a commercial 220 nm silicon-on-insulator (SOI) platform. Figure 3a shows an SEM image of the cascaded power splitter, with TE<sub>2</sub>-TE<sub>0</sub> mode splitter lengths of 18.5  $\mu$ m, 20  $\mu$ m, and 14.1  $\mu$ m, respectively. Three modes input from the left port, with higher-order modes decoupled in turn and divided equally between the upper and lower ports. Figure 3b presents an optical microscope image of the complete multimode 3-dB coupler, which is 127  $\mu$ m long with an adjustable width. Three modes enter Port  $I_1$ , propagate through the splitter, divide equally, and recombine via the MFI to form 3-dB couplers at Ports  $O_1$  and  $O_2$ . To evaluate insertion loss, straight waveguides with grating couplers were fabricated on the same chip (see “Methods”). Figure 3c–e display the normalized transmission spectra: at 1.55  $\mu$ m, the insertion losses for the TE<sub>2</sub>-TE<sub>0</sub> mode splitters are 0.2, 0.45, and 0.42 dB, respectively, with crosstalk below -20 dB. Across the C-band, maximum insertion losses reach 0.4, 0.67, and 0.76 dB, with all crosstalk values remaining below -17.3 dB. Experimental results align well with the simulated data provided in Supplementary Note 1. Slight output variations between Ports  $O_1$  and  $O_2$ , attributed to unavoidable sidewall etching roughness and linewidth variations during fabrication, remain within acceptable limits. The MKM<sup>40–43</sup> is the core structure of the multimode convolution accelerator. Although cascading two multimode 3-dB couplers with a phase shifter can achieve this function, that approach requires widening the phase-shifter’s bus waveguide—necessitating long tapers and increasing device size. To overcome this, we designed a compact MKM (Supplementary Note 3). Figure 3f shows an optical microscope image of the centrosymmetric device (325  $\times$  80  $\mu$ m<sup>2</sup>) with a central thermal insulation slot to prevent crosstalk. Three modes enter the left port, are split by cascaded mode splitters into three single-mode waveguides within the phase shifter, pass through another set of three cascaded splitters, and exit via the right port. The phase shifter uses three parallel waveguides (0.5  $\mu$ m spacing, 2  $\mu$ m width each) to avoid (de)multiplexing and long tapers and to reduce phase errors<sup>44,45</sup>. To ensure uniform temperature rise across all waveguides—avoiding excessive heating of the central waveguide—we use two parallel TiN microelectrodes (2  $\mu$ m wide, 140  $\mu$ m long, 0.1  $\mu$ m thick) with a spacing of  $G_{WE}$  = 1.92  $\mu$ m. Figure 3g plots the transmission spectra of the three modes at 1.55  $\mu$ m versus voltage. At the driving voltage of 3.6 V, all modes switch simultaneously, consuming 23.6 mW for a  $\pi$ -phase shift. Compared to the conventional approach using three separate microelectrodes that requires about 60 mW, our design reduces power consumption by approximately 60%<sup>46</sup>. Figure 3h–j show the normalized “on/off” spectra, with dashed lines marking the extinction ratios and solid lines indicating the inter-mode crosstalk (values below -30 dB omitted). Over the C-band, the insertion losses for TE<sub>2</sub>-TE<sub>0</sub> are 2.1, 1.3, and 1.5 dB, respectively. Maximum crosstalk is -13.2, -14.3, and -17.1 dB, and extinction ratios exceed 12 dB. The designed multimode

3-dB coupler and kernel modulator can be easily scaled to higher-order modes by cascading an additional higher-order mode splitter.

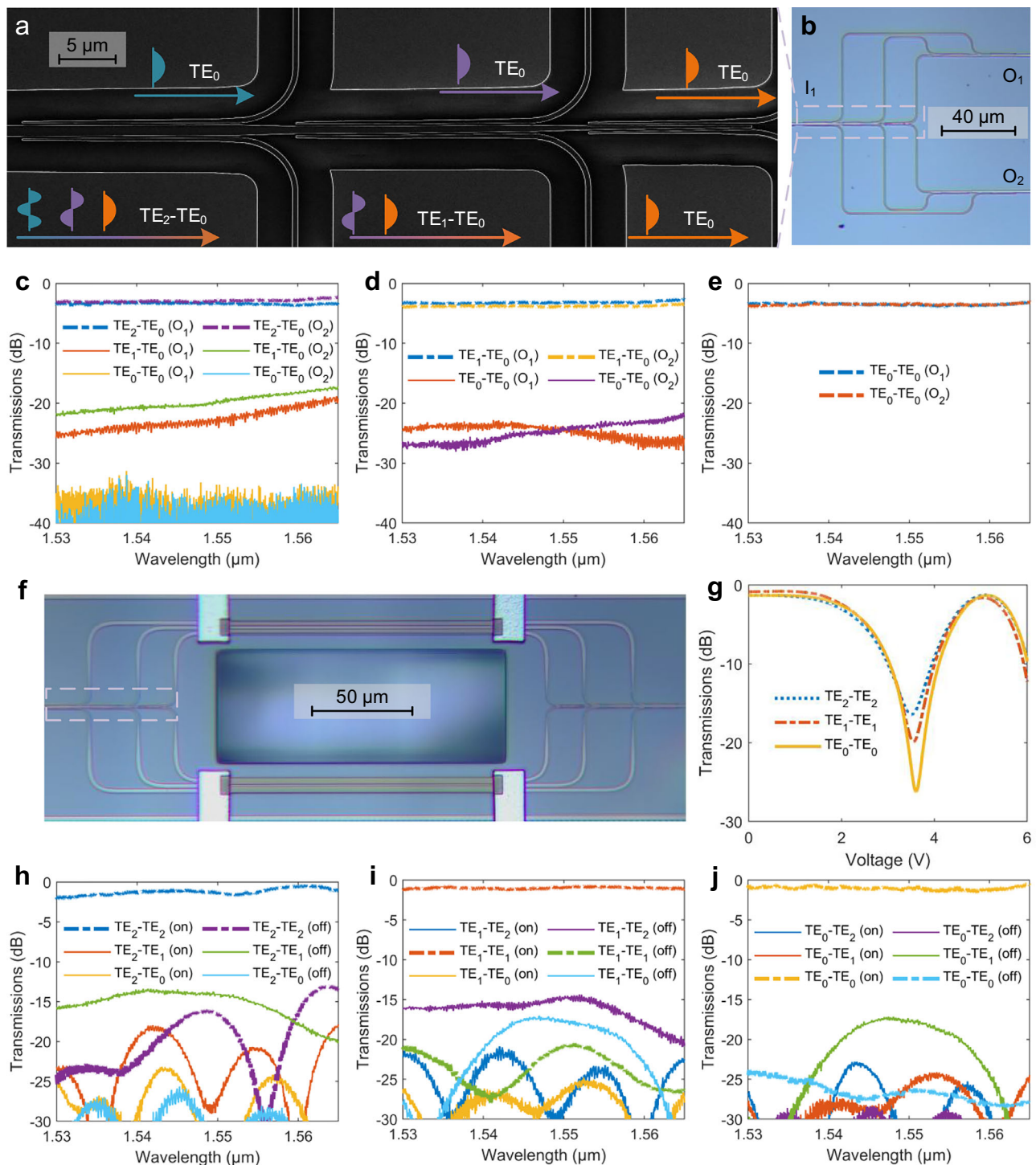
### Multimode convolution accelerator

Figure 4a displays the optical and electrical packaging of the chip, with vertical grating coupling and wire bonding completed. The chip, used as a proof of concept, measures 0.6  $\times$  0.7 mm<sup>2</sup>, and at 1.55  $\mu$ m, the total insertion losses of the photonic convolution accelerator are below 2.73 dB. The chip integrates three-mode 3-dB couplers, three-mode kernel modulators, and three-mode routing. A single-wavelength input generates three vectors, while four MKMs implement the loading of a 1  $\times$  4 convolution kernel. A microscope image of the fabricated photonic convolution accelerator chip, along with a schematic of the experimental apparatus, is provided in Supplementary Note 4. We tested 1000 sets of random input vectors and convolution kernels. Each convolution produced an output vector of length 6, yielding a total of 6000 data points. The input vector elements were random numbers between 0 and 1, and the convolution kernel elements ranged from 0 to 1/3, ensuring that the output values remained within 0 to 1. As shown in Fig. 4b, the experimentally obtained values align tightly along the diagonal, demonstrating precise computations with a high degree of reliability. The bit precision  $N_b$  is defined as:

$$N_b = \log_2 \left( \frac{\mu_{\max} - \mu_{\min}}{\sigma} \right) \quad (3)$$

where  $\mu_{\max}$  and  $\mu_{\min}$  are the maximum and minimum output values, respectively, and  $\sigma$  is the standard deviation of the error between the experimental and expected outputs. Gaussian fitting yields a normalized standard deviation of 0.0072, confirming a system computational precision of 7 bits. In Fig. 4c, we employ a CNN to perform ten-class classification of handwritten digits (0–9), with the convolution layer implemented via optical convolution (highlighted in orange). Images are first resized to 12  $\times$  12 pixels and flattened into a 144  $\times$  1 vector. The photonic convolution layer then uses four 4  $\times$  1 convolution kernels to generate four 146  $\times$  1 feature maps. To perform convolution on long vectors, the original operation is decomposed into multiple short-vector convolutions with subsequent encoding. We adopt the overlap-add method to compute the convolution; see Supplementary Note 5 for details. After applying the ReLU nonlinearity, the four 146  $\times$  1 feature maps are reshaped into a 584  $\times$  1 vector, which is then fed into a fully connected layer with 10 neurons using SoftMax activation. Although our architecture and experimental demonstrations are based on one-dimensional (1D) convolution, two-dimensional (2D) convolution can be equivalently portrayed as a 1D convolution by applying zero-padding to the kernel<sup>19</sup>. While the architecture does not inherently support convolution kernels with negative values, the kernel can be decomposed into two positive components and subtracted at the output to accommodate real-valued kernels. Offline training is carried out via stochastic gradient descent with back-propagation to minimize the cross-entropy loss (see Supplementary Note 6 for details). We performed 20 iterations of training on the MNIST dataset (70,000 images total; 60,000 for training, which was split with a training-to-validation ratio of 5:1). In the optical-domain test, 1000 images were loaded, yielding a classification accuracy of 95.2%, very close to the theoretical electrical-domain accuracy of 95.8%. The corresponding confusion matrices in Fig. 4d, e indicate that the effects of optical noise or distortion (due to higher-order dispersion) on network performance and effective precision are minimal.

To further demonstrate the capability of wavelength-parallel input, we utilized two different wavelengths to achieve recognition on the Fashion-MNIST dataset. As shown in Fig. 5a, b, using the same



**Fig. 3 | Design and experimental results of the multimode 3-dB coupler and the kernel modulator. a** SEM image of the cascaded mode splitter. **b** Optical microscope image of the multimode 3-dB coupler. **c–e** Measured transmission spectra in the C-band. Dashed lines indicate the mode corresponding to the mode splitter.

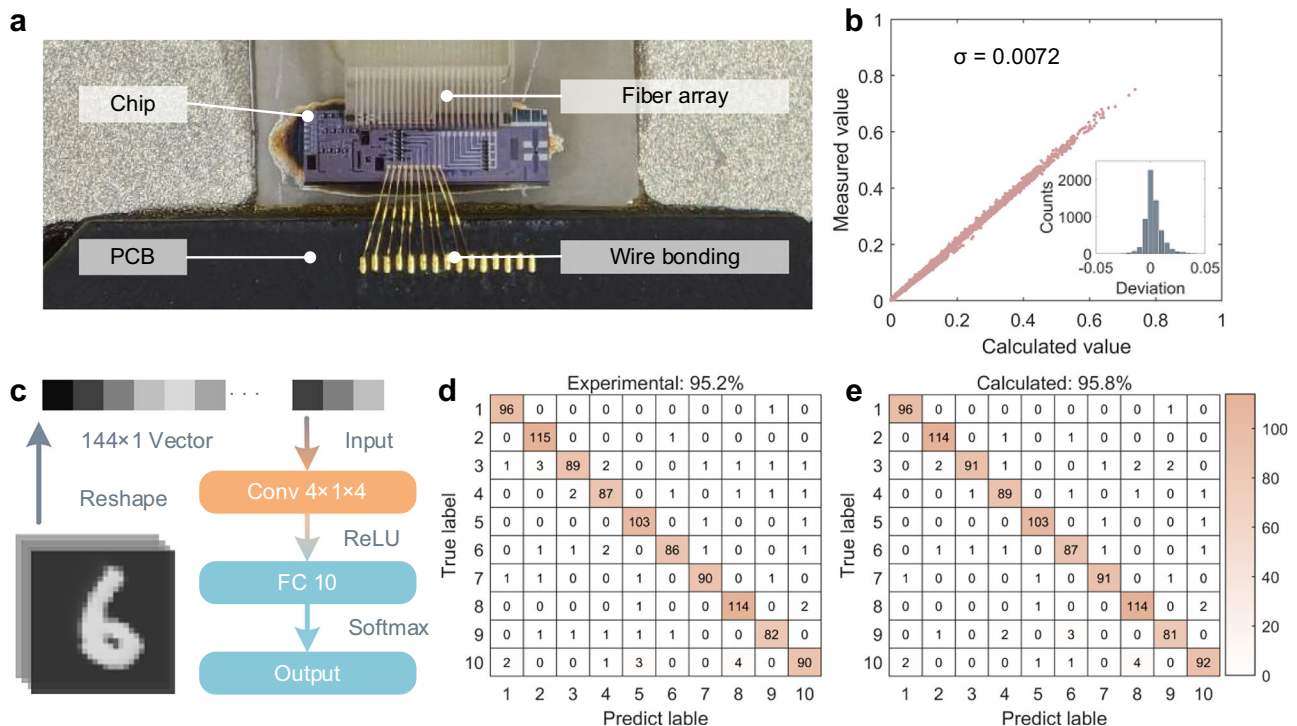
**f** Optical microscope image of the MKM. **g** Transmission spectra of three modes at 1550 nm under varying voltage. **h–j** Normalized transmission spectra in the “on” and “off” states. Dashed lines indicate variations in the same mode, whereas solid lines denote inter-mode crosstalk (values below  $-30$  dB are omitted).

precision verification method at wavelengths of 1530 nm and 1565 nm, the normalized standard deviations are 0.0094 and 0.0075, corresponding to computational precisions of 6 bits and 7 bits, respectively. After retraining the CNN with the same architecture as in Fig. 4c and loading the training parameters onto the chip, 500 Fashion-MNIST images were processed at each of the two wavelengths, achieving an accuracy of 87.9% (the confusion matrix is shown in Fig. 5c). In the electrical domain, the recognition accuracy was 89.2% (see Fig. 5d).

The slightly larger difference of 1.3% (compared to the previous 0.6% for MNIST) may be attributed to variations in precision at the different wavelengths.

## Discussion

The proposed multimode optical convolution accelerator, which integrates multimode 3-dB couplers, kernel modulators, and routing, has a compact footprint of  $0.6 \times 0.7 \text{ mm}^2$  and is compatible with



**Fig. 4 | Experimental results of the multimode convolution accelerator.**

**a** Optical and electrical packaging of the chip. **b** Scatter plot of measured convolution accuracy. Inset: Histogram of compute errors (differences between measured and computed values). **c** Schematic of the CNN architecture: orange regions

denote optically implemented convolutional layers. **d, e** Confusion matrices for recognition of 1000 MNIST test digits, comparing experimental results with computerized calculations.

130-nm commercial standard manufacturing processes. At 1.55  $\mu\text{m}$ , the total insertion losses of the photonic convolution accelerator are below 2.73 dB. Both the multimode 3-dB coupler and the kernel modulator were designed using inverse design, yielding high performance and a fabrication tolerance of  $\pm 15$  nm. Within the C-band, experimental results show that the multimode 3-dB coupler attains an extinction ratio exceeding 16 dB, while the MKM delivers an extinction ratio greater than 12 dB and requires only 23.6 mW for a  $\pi$ -phase shift. The convolution window's sliding and encoding rely entirely on passive optical components, minimizing power consumption. At 1.55  $\mu\text{m}$ , the accelerator provides a computational precision of 7 bits. In MNIST handwritten digit classification, optical-domain inference achieved an accuracy of 95.2%, closely matching the theoretical electrical-domain accuracy and validating the design's practical reliability. Additionally, wavelength-division multiplexing (WDM) enables parallel loading of multiple wavelengths, while mode-division multiplexing (MDM) provides distinct spatial modes in a single waveguide for efficient fan-in, both significantly boosting computational throughput. To verify the performance of WDM, tests at 1530 nm and 1565 nm yielded computational precisions of 6 bits and 7 bits, respectively, while recognition experiments on the Fashion-MNIST dataset recorded an optical-domain accuracy of 87.9%, showcasing the capability for complex tasks. Finally, the device provides a 4.38 THz available bandwidth (corresponding to the C-band) and achieves a compute density of up to 125.14 TOPS/ $\text{mm}^2$ , calculated as follows:

$$\frac{2 \times 3 \times 4 \times K \times \frac{4.38}{2K} \text{ THz}}{0.6 \times 0.7 \text{ mm}^2} \approx 125.14 \text{ TOPS/mm}^2 \quad (4)$$

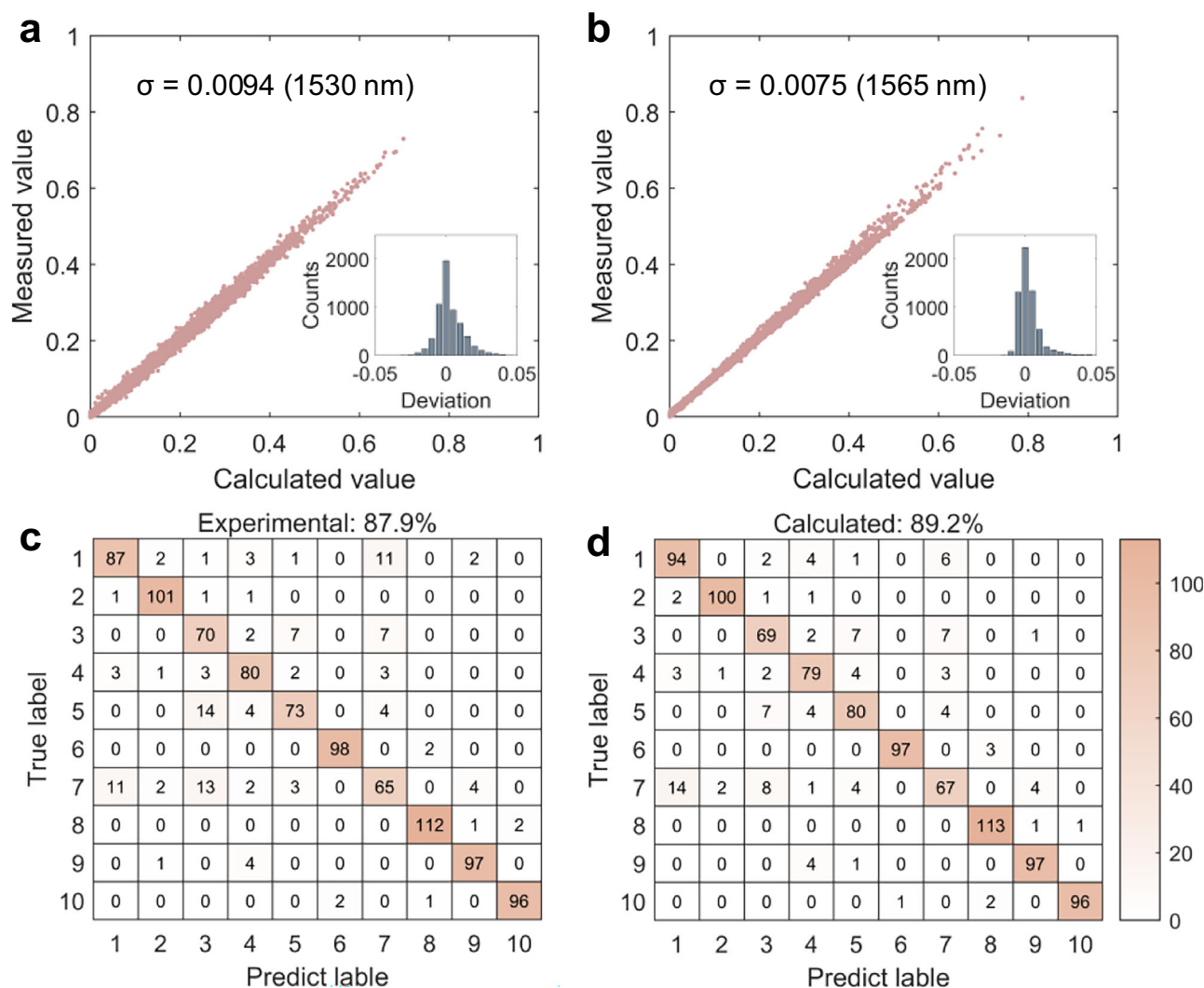
Here,  $2 \times 3 \times 4 \times K$  denotes the total MAC count per clock cycle across three modes, four kernels, and  $K$  wavelengths;  $4.38/K$  THz is the bandwidth of each wavelength channel, with modulation bandwidth

halved to prevent crosstalk (i.e.,  $4.38/(2K)$  THz); and  $0.6 \times 0.7 \text{ mm}^2$  is the chip footprint.

Our approach complements other innovative strategies recently demonstrated for photonic convolution accelerators. For instance, high-throughput convolvers have been realized using time-domain delay-and-interleave schemes<sup>22–24</sup>. Other works have ingeniously exploited synthetic dimensions<sup>19,20</sup>, demonstrating convolution in the frequency domain to unlock massive computational parallelism within a single modulated ring. Furthermore, redundancy-free architectures based on arrayed waveguide gratings (AWGs) have been developed to enhance computational efficiency<sup>47</sup>. For further comparisons of photonic convolution methods, see Supplementary Note 7. While these methods significantly advance the field, our work focuses specifically on overcoming the fan-in loss bottleneck through lossless mode-division multiplexing, offering a new pathway toward scalability in large-scale linear operations. While this work concentrates on the linear convolution step, integrating efficient on-chip nonlinear activation functions remains a critical next step for creating fully integrated photonic neural networks. Neurons based on nonlinear optical materials<sup>48</sup> and photodetector-driven modulators<sup>49–52</sup> are both highly promising solutions to this challenge.

In summary, our lossless mode-division fan-in strategy introduces a fundamentally new approach for large-scale on-chip optical summation by circumventing the insertion-loss bottleneck in single-mode systems. Through inverse-designed multimode components, we have demonstrated the viability of leveraging multiple orthogonal modes to achieve near-ideal fan-in efficiencies. Such architectures hold promising implications for photonic accelerators, where efficient fan-in plays a critical role in convolution and other linear operations. Moving forward, further optimization of device footprint, fabrication tolerance, and broad optical bandwidth will be key to integrating more modes and wavelengths for supporting energy-efficient, high-throughput photonic computing platforms.





**Fig. 5 | Experimental results of multi-wavelength parallel recognition on the Fashion-MNIST dataset. a, b** Scatter plots of convolution accuracy at 1530 nm and 1565 nm. Insets: Histograms of compute errors. **c, d** Confusion matrices comparing

experimental and computerized classification results for 1000 Fashion-MNIST test images.

## Methods

### Chip fabrication

The device was fabricated on a commercial 220 nm SOI wafer with a 2  $\mu$ m buried oxide. First, waveguide patterns were defined by deep ultraviolet (DUV) lithography and transferred into the silicon layer via dry etching. Next, for the Ge photodetectors, a localized ion implantation formed the bottom contact in the silicon waveguide, followed by heteroepitaxial growth of Ge and top-side doping. Afterward, an interlayer dielectric (ILD) was deposited and planarized via chemical mechanical polishing (CMP). Then, in the back-end-of-line (BEOL), contact vias were etched and filled with tungsten (W), and a TiN layer was subsequently deposited and patterned to form the on-chip microheaters for the thermo-optic phase shifters. Finally, multi-layer metal traces and vias were fabricated using a Cu single-Damascene process to interconnect the devices and route signals to the Al bond pads. The entire chip was then protected by a passivation layer, with windows opened over the bond pads for packaging and testing.

### Measurement methods

To facilitate the measurement of the normalized losses and crosstalk of the multimode 3-dB coupler and kernel modulator, mode (de) multiplexer connecting straight waveguides were fabricated on the

same chip. The experimental setup included a broadband light source, a vertical fiber-chip coupling stage, a spectrum analyzer, and a power meter. Prior to vertical coupling, a polarization controller was used to select the TE polarization state. Grating couplers connected single-mode fibers to the silicon waveguides, each incurring a measured loss of approximately 4 dB. The output light was divided in a 1:9 ratio—10 % was sent to a high-sensitivity power meter for continuous, real-time monitoring of the coupling efficiency, and 90 % to the spectrum analyzer to ensure sufficient optical power for an adequate signal-to-noise ratio. All measurements were conducted at room temperature.

### Data availability

The data are available as part of the Supplementary files. Source data are provided with this paper.

### References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Shastri, B. J. et al. Photonics for artificial intelligence and neuro-morphic computing. *Nat. Photon.* **15**, 102–114 (2021).
3. Zhou, H. et al. Photonic matrix multiplication lights up photonic accelerator and beyond. *Light Sci. Appl.* **11**, 30 (2022).

4. Xu, H., Dai, D. & Shi, Y. Silicon integrated nanophotonic devices for on-chip multi-mode interconnects. *Appl. Sci.* **10**, 6365 (2020).
5. Goossens, J.-W., Yousefi, M. I., Jaouën, Y. & Hafermann, H. Polarization-division multiplexing based on the nonlinear Fourier transform. *Opt. Express* **25**, 26437–26452 (2017).
6. Yan, Y. et al. High-capacity millimetre-wave communications with orbital angular momentum multiplexing. *Nat. Commun.* **5**, 4876 (2014).
7. Goodman, J. W., Dias, A. R. & Woody, L. M. Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Opt. Lett.* **2**, 1–3 (1978).
8. Cheng, J. et al. Multimodal deep learning using on-chip diffractive optics with in situ training capability. *Nat. Commun.* **15**, 6189 (2024).
9. Goodman, J. W. Fan-in and fan-out with optical interconnections. *Opt. Acta. Int. J. Opt.* **32**, 1489–1496 (1985).
10. Chen, J., Clark, L. T. & Cao, Y. Maximum fan-in/out. *IEEE Circuits Devices Mag.* **21**, 12–20 (2005).
11. Ji, L. & Heuring, V. P. Impact of gate fan-in and fan-out limits on optoelectronic digital circuits. *Appl. Opt.* **36**, 3927–3940 (1997).
12. Bernstein, L. et al. Freely scalable and reconfigurable optical hardware for deep learning. *Sci. Rep.* **11**, 3144 (2021).
13. Ma, G., Yu, J., Zhu, R. & Zhou, C. Optical multi-imaging-casting accelerator for fully parallel universal convolution computing. *Photon. Res.* **11**, 299–312 (2023).
14. Wang, T. et al. Image sensing with multilayer nonlinear optical neural networks. *Nat. Photon* **17**, 408–415 (2023).
15. Crespi, A. et al. Integrated multimode interferometers with arbitrary designs for photonic boson sampling. *Nat. Photon.* **7**, 545–549 (2013).
16. Marcattili, E. A. J. Dielectric rectangular waveguide and directional coupler for integrated optics. *Bell Syst. Tech. J.* **48**, 2071–2102 (1969).
17. Jalas, D. et al. What is—and what is not—an optical isolator. *Nat. Photon.* **7**, 579–582 (2013).
18. Xu, S., Wang, J., Yi, S. & Zou, W. High-order tensor flow processing using integrated photonic circuits. *Nat. Commun.* **13**, 7820 (2022).
19. Fan, L. et al. Multidimensional convolution operation with synthetic frequency dimensions in photonics. *Phys. Rev. Appl.* **18**, 034088 (2022).
20. Fan, L., Wang, K., Wang, H., Dutt, A. & Fan, S. Experimental realization of convolution processing in photonic synthetic frequency dimensions. *Sci. Adv.* **9**, eadi4956 (2023).
21. Meng, X. et al. Compact optical convolution processing unit based on multimode interference. *Nat. Commun.* **14**, 2779 (2023).
22. Xu, X. et al. Photonic perceptron based on a Kerr microcomb for high-speed, scalable, optical neural networks. *Laser Photon. Rev.* **14**, 2000070 (2020).
23. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
24. Bai, B. et al. Microcomb-based integrated photonic processing unit. *Nat. Commun.* **14**, 203 (2023).
25. Chen, G., Yu, Y. & Zhang, X. Monolithically mode division multiplexing photonic integrated circuit for large-capacity optical interconnection. *Opt. Lett.* **41**, 3543–3546 (2016).
26. Dai, D., Wang, J., Chen, S., Wang, S. & He, S. Monolithically integrated 64-channel silicon hybrid demultiplexer enabling simultaneous wavelength- and mode-division-multiplexing. *Laser Photon. Rev.* **9**, 339–344 (2015).
27. Du, J. et al. Mode division multiplexing: from photonic integration to optical fiber transmission. *Chin. Opt. Lett.* **19**, 091301 (2021).
28. Jain, M., Saihpal, V., Singh, N. & Singh, S. B. An overview of variants and advancements of PSO algorithm. *Appl. Sci.* **12**, 8392 (2022).
29. Liu, Z. et al. Integrated nanophotonic wavelength router based on an intelligent algorithm. *Optica* **6**, 1367–1373 (2019).
30. Kirkpatrick, S., Gelatt, C. D. Jr. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
31. Molesky, S. et al. Inverse design in nanophotonics. *Nat. Photon* **12**, 659–670 (2018).
32. Nikkhah, V. et al. Inverse-designed low-index-contrast structures on a silicon photonics platform for vector-matrix multiplication. *Nat. Photon* **18**, 501–508 (2024).
33. Yang, K. Y. et al. Inverse-designed non-reciprocal pulse router for chip-based LiDAR. *Nat. Photon* **14**, 369–374 (2020).
34. Sun, S. et al. Inverse design of ultra-compact multimode waveguide bends based on the free-form curves. *Laser Photon. Rev.* **15**, 2100162 (2021).
35. Sun, S. et al. Ultra-compact multimode waveguide routings based on the free-form curves. *J. Lightwave Technol.* **43**, 764–775 (2025).
36. Xu, H. & Shi, Y. Ultra-broadband dual-mode 3-dB power splitter based on a Y-junction assisted with mode converters. *Opt. Lett.* **41**, 5047–5050 (2016).
37. Zhong, Z. et al. T-branch waveguide mirror for multimode optical splitter with arbitrary power ratios. *IEEE J. Quantum Electron.* **57**, 8400106 (2021).
38. Xu, J., Liu, Y., Guo, X., Song, Q. & Xu, K. Inverse design of a dual-mode 3-dB optical power splitter with a 445 nm bandwidth. *Opt. Express* **30**, 26266–26274 (2022).
39. Chang, W. et al. Inverse design and demonstration of an ultra-compact broadband dual-mode 3 dB power splitter. *Opt. Express* **26**, 24135–24144 (2018).
40. Sun, S. et al. Compact, scalable, fast-response multimode 2 × 2 optical switch based on inverse design. *Laser Photon. Rev.* **18**, 2400213 (2024).
41. Sun, S. et al. Large-scale high-density multimode optical switch matrix. *Laser Photon. Rev.* **19**, 2400930 (2025).
42. Sun, C. et al. Key multimode silicon photonic devices inspired by geometrical optics. *ACS Photon.* **7**, 2037–2045 (2020).
43. Zhao, W. et al. High-performance silicon polarization switch based on a Mach-Zehnder interferometer integrated with polarization-dependent mode converters. *Nanophotonics* **11**, 2293–2301 (2022).
44. Song, L. et al. Low-loss calibration-free 2 × 2 Mach-Zehnder switches with varied-width multimode-interference couplers. *J. Lightwave Technol.* **40**, 5254–5259 (2022).
45. Song, L. et al. Toward calibration-free Mach-Zehnder switches for next-generation silicon photonics. *Photon. Res.* **10**, A54–A61 (2022).
46. Liu, S. et al. Comparison of thermo-optic phase-shifters implemented on CUMEC silicon photonics platform. in *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, Vol. 12059, 111–116 (SPIE, 2021).
47. Zhang, S. et al. Redundancy-free integrated optical convolver for optical neural networks based on arrayed waveguide grating. *Nanophotonics* **13**, 19–28 (2024).
48. Chen, C. et al. Ultra-broadband all-optical nonlinear activation function enabled by MoTe<sub>2</sub>/optical waveguide integrated devices. *Nat. Commun.* **15**, 9047 (2024).
49. Pour Fard, M. M. et al. Experimental realization of arbitrary activation functions for optical neural networks. *Opt. Express* **28**, 12138–12148 (2020).
50. Williamson, I. A. D. et al. Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *IEEE J. Sel. Top. Quantum Electron.* **26**, 7800212 (2020).
51. Huang, C. et al. A silicon photonic-electronic neural network for fibre nonlinearity compensation. *Nat. Electron* **4**, 837–844 (2021).
52. Bandyopadhyay, S. et al. Single-chip photonic deep neural network with forward-only training. *Nat. Photon.* **18**, 1335–1343 (2024).

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (62425504 to J.D., U21A20511 to J.D.).



## Author contributions

S. S. and S. Z. conceived the idea. S. S. designed the silicon photonic chip. S. Z. conducted the theoretical analysis. S. S. and S. Z. conducted the experiments and analyzed the data. B.W., S.J. and H.Z. involved in the discussion. S. S. wrote the original manuscript with S. Z.'s contributions. S. S., S. Z., B.W., S.J., B.Z., and J.D. revised and edited the manuscript. J.D. and X.Z. supervised and coordinated all the work.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-62954-w>.

**Correspondence** and requests for materials should be addressed to Jianji Dong.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025