





# Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network

Changming Wu<sup>1</sup>, Heshan Yu<sup>2</sup>, Seokhyeong Lee<sup>1</sup>, Ruoming Peng<sup>1</sup>, Ichiro Takeuchi <sup>2</sup> & Mo Li <sup>1,3</sup> 

Neuromorphic photonics has recently emerged as a promising hardware accelerator, with significant potential speed and energy advantages over digital electronics for machine learning algorithms, such as neural networks of various types. Integrated photonic networks are particularly powerful in performing analog computing of matrix-vector multiplication (MVM) as they afford unparalleled speed and bandwidth density for data transmission. Incorporating nonvolatile phase-change materials in integrated photonic devices enables indispensable programming and in-memory computing capabilities for on-chip optical computing. Here, we demonstrate a multimode photonic computing core consisting of an array of programmable 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  $\text{Ge}_2\text{Sb}_2\text{Te}_5$  during phase transition to control the waveguide spatial modes with a very high precision of up to 64 levels in modal contrast. This contrast is used to represent the matrix elements, with 6-bit resolution and both positive and negative values, to perform MVM computation in neural network algorithms. 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 toward large-scale photonic neural networks with ultrahigh computation throughputs.

<sup>1</sup>Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195, USA. <sup>2</sup>Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742, USA. <sup>3</sup>Department of Physics, University of Washington, Seattle, WA 98195, USA. email: [moli96@uw.edu](mailto:moli96@uw.edu)

The unmet gap between the rate of energy efficiency improvement of current digital electronics and the fast-growing load of computation by emerging applications such as machine learning and artificial intelligence<sup>1,2</sup> has once again brought optical computing into focus<sup>3–6</sup>. Integrated photonics provides a scalable hardware platform to realize large-scale optical networks on a chip, which affords an enormous bandwidth density that is unreachable for electronics<sup>7–9</sup>. To use integrated photonics for optical computing, programmable photonic components and nonlinear elements are indispensable building blocks. Phase-change materials (PCM) recently emerged as an ideal material system to realize optical programmability<sup>10–12</sup>. 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, lens, and metamaterials to be demonstrated<sup>13–21</sup>. The phase change in the chalcogenide family of Ge-Sb-Te alloys is nonvolatile, requiring no sustaining power supply to retain the programmed state or stored information<sup>19–26</sup>. Their use in programmable photonic devices thus can have a significant advantage in power consumption over electro-optic<sup>27–29</sup> or thermo-optic methods<sup>30–32</sup>. Photonic devices incorporating those nonvolatile PCMs thus can realize optical memories and perform in-memory computing simply by measuring the transmission of the optical input data through the programmed device<sup>33–35</sup>. Proliferating these phase-change photonic devices in a scalable network, prototypes of optical neural networks (ONN) have been proposed and demonstrated<sup>35–38</sup>.

Here, we report a programmable waveguide mode converter based on a phase-gradient metasurface made of phase-change material Ge<sub>2</sub>Sb<sub>2</sub>Te<sub>5</sub> (GST). 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<sub>0</sub> and TE<sub>1</sub> 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 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.

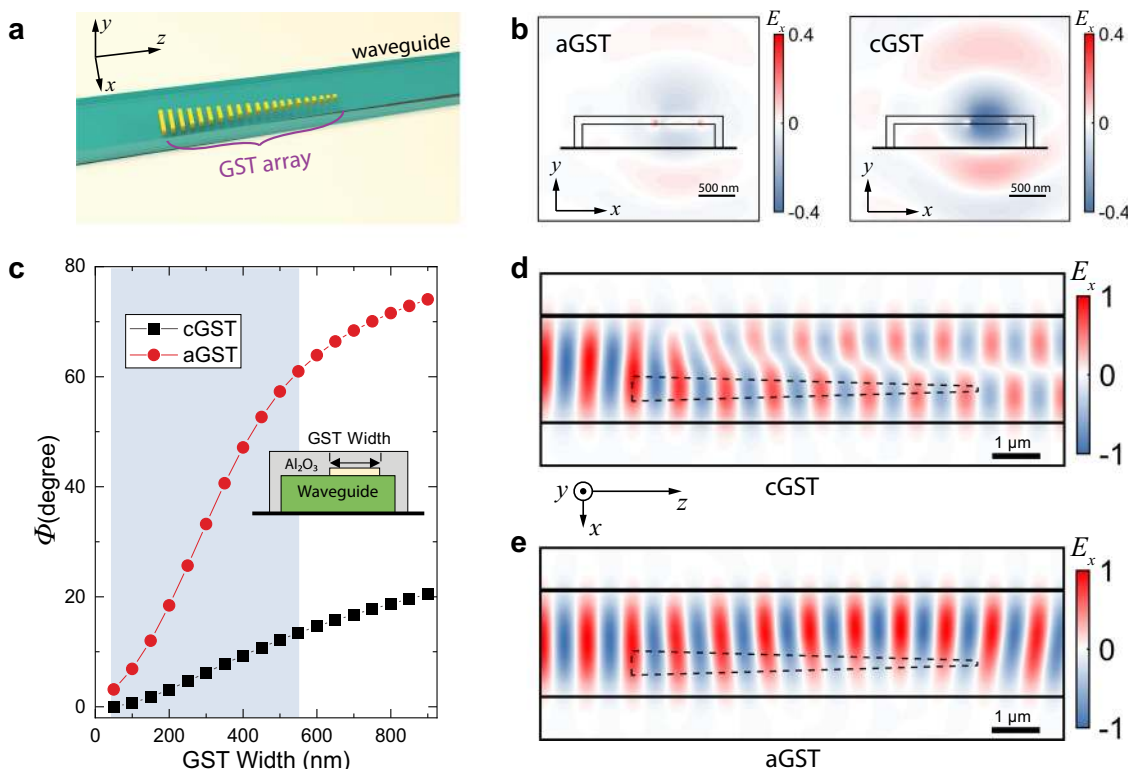
## Results

### High-precision programmable phase-change mode converter.

The design of the PMMC is based on the principle of a phase-gradient metasurface but replacing noble metals with phase-change materials<sup>39</sup>. Fig. 1a shows a 3D schematic of the design, which consists of a linear array of GST nano-antennae directly integrated on a silicon nitride (SiN) waveguide. Each GST nano-antenna scatters the waveguide mode and causes a phase shift  $\Phi$ , which depends on its geometry (e.g., width), as well as the refractive index of its material (Fig. 2b). 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:  $k_{\text{mode}1} - k_{\text{mode}2}$ , it satisfies the phase-matching condition and facilitates the conversion between the two modes. Such phase-gradient metasurfaces for waveguide mode conversion were realized using noble metals or dielectrics materials and 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 (aGST), its refractive index  $n$  is  $\sim 4.7$  (representative value in the literature, the same hereafter)<sup>40</sup>. In contrast, when it is turned to the crystalline phase (cGST),  $n$  increases to  $\sim 7.5$  with a drastic change of 2.8 over the whole measured spectral range from 1540 nm to 1580 nm (See Supplementary Fig. 1a for more detailed information). This change will significantly modify the scattered phase of each GST nano-antenna (Fig. 2b) so as to modify the metasurface's function. Fig. 1c 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. Since cGST has a much larger  $n$ , the scattered phase shows a much stronger dependence on the width than the aGST phase. 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 (see Supplementary Note 2 and 3 for details). The entire metasurface consists of an array of 25 nano-antennae with tapering widths from 510 nm to 84 nm (shaded region in Fig. 1c) and is patterned on a SiN waveguide 1.8  $\mu\text{m}$  wide and 330 nm thick. The waveguide supports two transverse-electric modes: the fundamental TE<sub>0</sub> mode and the first-order TE<sub>1</sub> mode. We design the metasurface, in the cGST phase, to have a uniform  $d\Phi = 2.5^\circ$  for every  $dx = 400$  nm to satisfy the generalized phase-matching condition,  $k_0(n_{\text{TE}0} - n_{\text{TE}1}) = N \cdot d\Phi/dx$ , where  $k_0$  is the free-space wavevector,  $n_{\text{TE}0}$  and  $n_{\text{TE}1}$  are the effective index of the TE<sub>0</sub> and TE<sub>1</sub> modes, respectively, and  $N$  is the number of interactions between the guided modes and the metasurface. The cGST metasurface thus can efficiently convert the TE<sub>0</sub> mode to the TE<sub>1</sub> mode, as shown by the finite-difference time-domain (FDTD) simulation result in Fig. 1d. When the GST is transitioned to the aGST phase, as shown in Fig. 1c, the  $d\Phi/dx$  is much reduced and thus insufficient for the phase-matching condition so that mode conversion between TE<sub>0</sub> and TE<sub>1</sub> modes does not occur, which is clearly seen in Fig. 1e. Therefore, the GST phase-gradient metasurface, as designed here, functions as a programmable waveguide mode converter controlled by the tunable material phase of the GST.

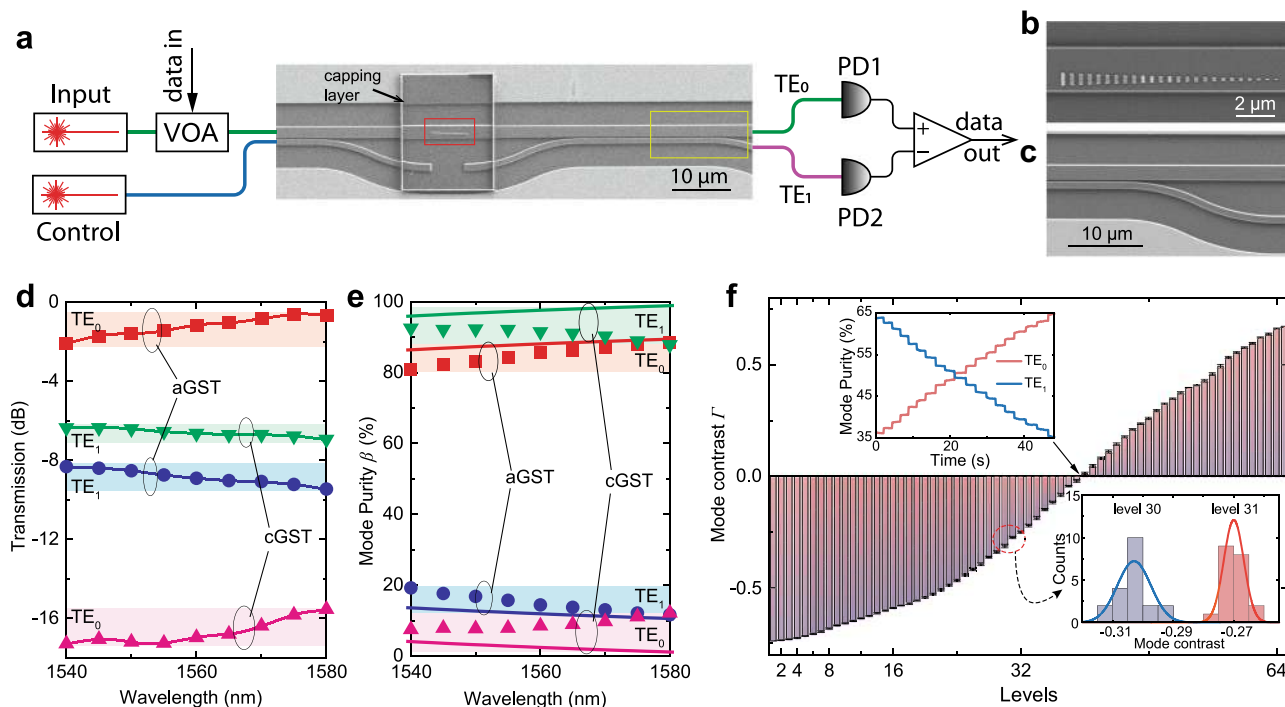
Fig. 2a–c shows the scanning electron microscope images of the complete PMMC device. The 30 nm thick GST film is deposited by sputtering on Si<sub>3</sub>N<sub>4</sub> on an oxidized silicon substrate. It is then patterned into metasurface with electron beam lithography and plasma etching, and conformally encapsulated with a 218-nm-thick layer of Al<sub>2</sub>O<sub>3</sub> deposited by atomic layer deposition. The photonic circuits of Si<sub>3</sub>N<sub>4</sub>, including multimode waveguides, directional couplers and grating couplers, are patterned with standard processes<sup>17</sup>. A pair of asymmetric directional couplers (Fig. 2c) is designed to function as mode selectors to selectively couple only the TE<sub>1</sub> mode component in the multimode waveguide with the TE<sub>0</sub> mode component in the single-mode waveguide (See Supplementary Note 5 for details). Fig. 2a depicts the measurement and control scheme. To program the PMMC, we use optical pulses to control the phase of the GST film for simplicity<sup>41</sup>. Previously, electrical control using integrated micro-heaters has been demonstrated by a number of groups, including us<sup>17,26,42–44</sup>. When operating the PMMC, an optical signal is input in the TE<sub>0</sub> mode to the PMMC and converted to TE<sub>1</sub> mode with a proportion controlled by the state of the GST metasurface. At the output of the PMMC, the TE<sub>1</sub> component is separated by the mode selector and coupled out at the second port while the TE<sub>0</sub> component remains in and outputs from the multimode waveguide. The output powers of both modes are measured to determine their respective transmission coefficients. Fig. 2d shows the transmission spectrum of the PMMC when the metasurface is set to be either in the 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<sub>0</sub> mode with a high transmission  $T^{\text{on}}$  over a broad wavelength range (1540–1580 nm).



**Fig. 1** Design of the phase-gradient metasurface mode converter. **a** 3D illustration of the devices. **b** FDTD simulation of the scattered electric field by one nano-antenna when the GST is in aGST (left panel) and cGST (right panel) phases, respectively, showing the distinctive difference. **c**. 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: cross-sectional view of the structure. **d, e** FDTD simulation results showing effective mode conversion from the TE<sub>0</sub> mode to the TE<sub>1</sub> mode when the GST is in crystalline phase (**d**), but only a small perturbation when the GST is in amorphous phase (**e**).

The lowest insertion loss is 0.9 dB at 1575 nm wavelength. A small portion (<−10 dB) of the TE<sub>1</sub> 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 TE<sub>0</sub> mode to the TE<sub>1</sub> mode effectively. In this off-state for the TE<sub>0</sub> mode, its transmission  $T^{\text{off}}$  is <−15 dB over the entire measured bandwidth. The corresponding switching extinction ratio, defined as  $\Delta T/T^{\text{off}} = (T^{\text{on}} - T^{\text{off}})/T^{\text{off}}$ , is ~16 dB or 4000%, which is more than 10-fold improvement compared to previously reported switch devices using GST<sup>24,43,45</sup>. 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<sup>2</sup>, 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<sub>1</sub> mode is switched in the opposite way to the TE<sub>0</sub> mode. From aGST to cGST phase, the TE<sub>1</sub> 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 in the multimode waveguide, defined as  $\beta_{\text{TE}_0(\text{TE}_1)} = P_{\text{TE}_0(\text{TE}_1)}/(P_{\text{TE}_0} + P_{\text{TE}_1})$ , where  $P_{\text{TE}_0}$  ( $P_{\text{TE}_1}$ ) is the power in the TE<sub>0</sub> (TE<sub>1</sub>) mode. The PMMC shows very high performance in controlling mode purity. As shown in Fig. 2e, when switching the GST from aGST to cGST phase, the PMMC efficiently converts TE<sub>0</sub> mode to TE<sub>1</sub> mode, changing the mode purity from  $\beta_{\text{TE}_0} > 80\%$  to  $\beta_{\text{TE}_1} > 85\%$  over a broad bandwidth, showing an excellent agreement with the numerical simulation results.

**PMMC photonic kernel.** 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<sub>1</sub> mode purity  $\beta_{\text{TE}_1}$  increases stepwise. Since the mode selector separates the two modes, we can measure their power and calculate the difference to determine the mode contrast  $\Gamma = \beta_{\text{TE}_0} - \beta_{\text{TE}_1}$ , which is used as a programming parameter. Fig. 2f demonstrates the multi-level programmability of the PMMC, in which  $\Gamma$  is sequentially set to 64 distinguishable levels between −0.73 to +0.67 at 1555 nm. Since the theoretical range of  $\Gamma$  is (−1, 1), it is an ideal parameter to represent the elements in the matrix  $w$ , with both positive and negative values, in a 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  $\Gamma$  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 Fig. 2f 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 clearly demonstrate the device’s programming resolution and accuracy. The demonstrated 64-level programmability of the PMMC—the highest to the best of our knowledge for phase-change photonic devices<sup>24</sup>—corresponds to 6-bit resolution in setting  $w$ , which is critical to the training and inference precision of the neural network<sup>46,47</sup>.



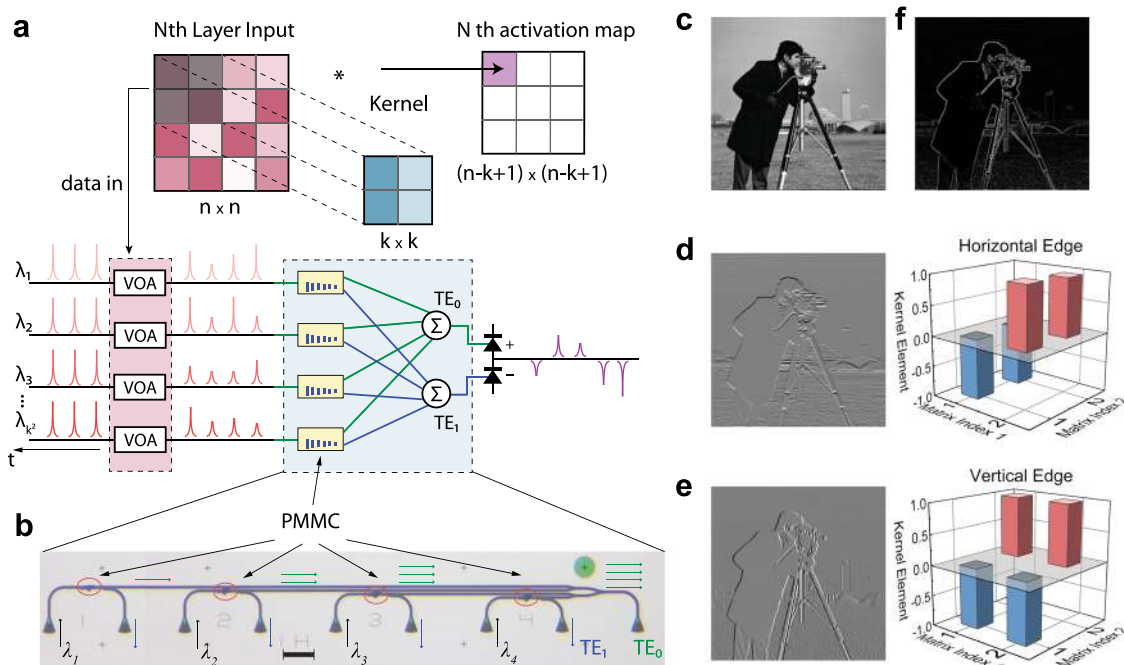
**Fig. 2 Operation of the programmable metasurface mode converter (PMMC).** **a** Scanning electron microscope (SEM) image of the complete device and the measurement and control schematics. The complete PMMC device consists of an encapsulated GST phase gradient metasurface (red box) and a mode selector (yellow box). The white box appears because of the edge of the 218-nm-thick  $\text{Al}_2\text{O}_3$  encapsulating layer. **b** Zoomed-in SEM image of the phase-gradient metasurface on the waveguide before depositing the  $\text{Al}_2\text{O}_3$  encapsulation layer for better imaging. **c** Zoomed-in SEM image of the  $\text{TE}_0/\text{TE}_1$  mode selector. **d** The transmission coefficient (insertion loss) of the devices for  $\text{TE}_0$  and  $\text{TE}_1$  modes and aGST and cGST phases. The transmission of the  $\text{TE}_0$  mode is switched with a high extinction ratio of >16 dB or 4000%. **e** The mode purity is controlled by the mode converter to >80% for both modes. **f** The programmable mode converter controls the mode contrast  $\Gamma$  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 at two adjacent levels (level 30 and 31). The well-separated histograms demonstrate the programming repeatability and accuracy.

We harness the PMMC's high-precision programmability and in-memory computing capability to demonstrate an optical convolutional neural network (OCNN)<sup>28–30,48</sup>. 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 OCNN using a small network of PMMCs to implement patch-kernel matrix multiplication to compute convolution. Fig. 3a 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 Fig. 3a) 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 grayscale image. The corresponding element  $W_{ij}$  of the kernel matrix is programmed as the mode contrast  $\Gamma$  of each PMMC. The resulting transmitted power of  $\text{TE}_0$  and  $\text{TE}_1$  modes are then summed incoherently using two photodetectors. Their output 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  $\Gamma$  of our PMMCs can assume both positive and negative values, it can represent the kernel matrix elements without the need of an additional offset, which otherwise would take additional steps to set in each computation cycle.

**Convolutional edge detection with PMMC core.** Experimentally, we build a small-scale, four-channel system with four PMMCs to represent a  $2 \times 2$  kernel matrix, as shown in the optical image in Fig. 3b. As a demonstration, we perform the convolution of a  $256 \times 256$  8-bit grayscale image of a cameraman (Fig. 3c) to detect its edge features. As shown in Fig. 3b, the  $\text{TE}_0$  mode output coming from all the PMMCs is combined using on-chip Y-junctions, while the  $\text{TE}_1$  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  $\text{TE}_0$  mode power by this factor when calculating the power differences between two



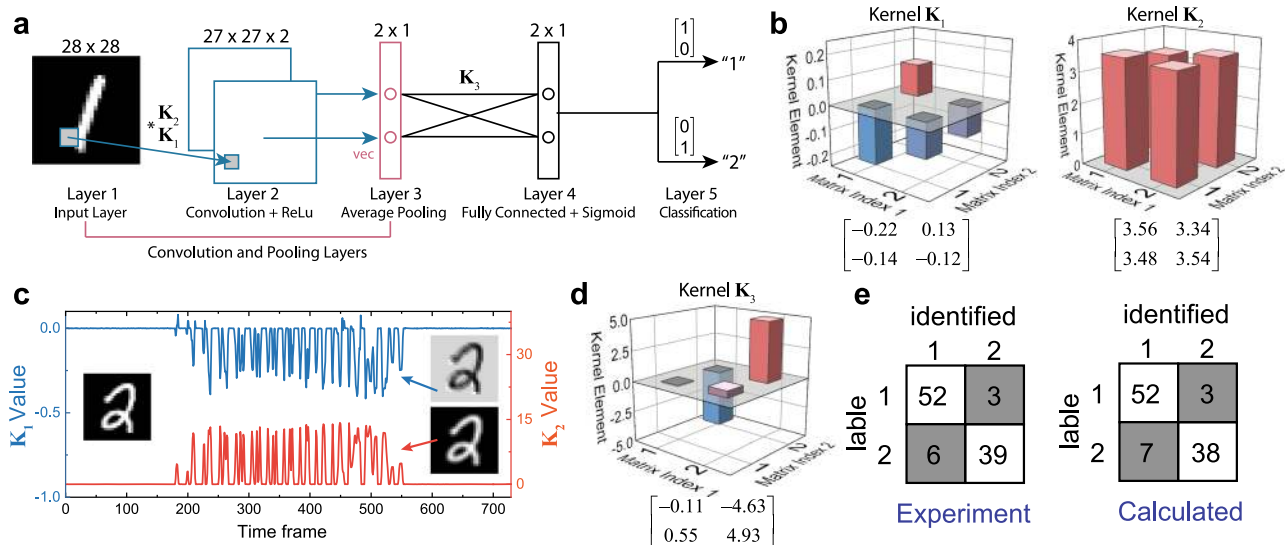
**Fig. 3** Using a PMMC array as a photonic computing core for convolutional image processing. **a** Schematic of optical convolution for image processing. An array of  $k^2$  PMMC is programmed to store the kernel matrix. A patch of pixels of an image is encoded as optical pulses and input into  $k^2$  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, 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.

modes. To detect vertical and horizontal edges, kernel matrices as in the right column of Fig. 3d, e are used, and so are the PMMCs programmed. Take the 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  $i, j$  are the indices of the input image matrix. Each kernel element  $W_{ij}$  is stored as the mode contrast value  $I$  in the corresponding PMMC, with  $W_{ij} = 1$  ( $-1$ ) corresponding to the fully aGST (cGST) phase (see Supplementary Note 9 for a more detailed description about the operation procedure). The computed images after convolution without any post-processing are shown in the left column of Fig. 3d, e, for horizontal and vertical edge detection, respectively. The two images are then added to produce the right image in Fig. 3b, 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 agrees very well with the calculated result using conventional image processing algorithms (see Supplementary Fig. 16). This result verifies the capacity and fidelity of optical convolution performed with the PMMC-based photonic kernel, which is a prerequisite for an OCNN.

**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 the 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. Fig. 4a illustrates the architecture and processes of the OCNN. The  $28 \times 28$  pixels, 8-bit grayscale images of number “1” or “2” are fed into the input layer as optical signals. The data is then convolved with two  $2 \times 2$  photonic kernels  $K_1$  and  $K_2$  to generate two  $27 \times 27$  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 a  $2 \times 1$  vector. This vector is then fed into the fully connected layer with a  $2 \times 2$  photonic weight bank  $K_3$  programmed in the PMMC array, added with a bias  $b_2$  and applied the standard sigmoid function. The final output is a vector that gives the identified class of the input image, that is,  $[1 \ 0]^T$  corresponds to the number “1” and  $[0 \ 1]^T$  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.

Before using the OCNN, we first train all the parameters in the layers with the standard back-propagation algorithm using the gradient descent method<sup>49</sup>. The training set consists of 11,000 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 Fig. 4b, d. We then program the PMMC array to represent these elements. In Fig. 4c, we show the raw data of the convolutional activation maps encoded in time series of optical signals, which are the output from the PMMC array after the input image convolves with the photonic kernel  $K_1$  and  $K_2$ . Since each photonic processing layer results in electrical signals



**Fig. 4** Building an optical CNN for imaging recognition. **a** Operation procedure of using the OCNN to recognize handwriting numbers from the MNIST database. The OCNN consists of a convolution layer with two kernels, a pooling and a fully connected layer. The output gives the answer whether the input image is "1" or "2". **b** The convolution kernel matrices  $K_1$  and  $K_2$  generated by training the OCNN. **c** Raw output data of the convolution layer of two kernel matrices. **d** The weight bank matrix used in the fully connected layer. **e** The recognition results from the experiment with the OCNN (left) and the calculation with a computer (right) show an excellent agreement.

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. Further experimental details are included in the Supplementary Note 9. 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. Fig. 4e 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.

## Discussion

In summary, we have demonstrated a compact programmable waveguide mode converter using 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  $\sim 1$  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<sup>50,51</sup>. Using wavelength division multiplexing (WDM) can further increase the number of parallel computation. The  $2 \times 2$  array prototype system demonstrated in this work performs optical computation incoherently in a broadband. It can be scaled up toward a large network using a photonic crossbar array

architecture<sup>52–57</sup> (see Supplementary Note 10 for details of such a design), and compares favorably with other photonic computing schemes using coherent methods<sup>30</sup> or optical resonators<sup>28,58,59</sup>. The feasible size ( $n \times m$ ) of such crossbar arrays will not be limited by the insertion loss of the PMMC ( $\sim 7$  dB for TE<sub>1</sub> mode, Fig. 2d); 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  $\sim 80 \times 20 \mu\text{m}^2$  (Fig. 2a, including the mode selector). 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 upperbound value of 25 TOPS/mm<sup>2</sup> (Tera-operations per second per mm<sup>2</sup>), which is significantly higher than that of digital electronic accelerators such as GPUs and tensor processing units (TPUs)<sup>60,61</sup>. Using silicon instead of silicon nitride can further reduce the device footprint to increase the computing density<sup>62</sup>. 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<sup>27,28,63</sup>. Alternatively, a hybrid photonic-electronic system may optimally balance energy-efficiency and speed advantages of photonic systems, while realizing flexible non-linearity, connectivity, and training precision using microelectronics<sup>26,64,65</sup>. 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.

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Code availability

No custom computer code or mathematical algorithm is used to generate the results that are reported in this study.

Received: 16 April 2020; Accepted: 25 November 2020;

Published online: 04 January 2021

## References

- Marr, B., Degnan, B., Hasler, P. & Anderson, D. Scaling Energy Per Operation via an Asynchronous Pipeline. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. **21**, 147–151 (IEEE, 2013).
- Jones, N. How to stop data centres from gobbling up the world's electricity. *Nature* **561**, 163–166 (2018).
- Athale, R. & Psaltis, D. Optical computing: past and future. *Opt. Photon. News* **27**, 32–39 (2016).
- Solli, D. R. & Jalali, B. Analog optical computing. *Nat. Photon.* **9**, 704–706 (2015).
- Prucnal, P. R. & Shastri, B. J. *Neuromorphic Photonics* (CRC Press, 2017).
- Caulfield, H. J. & Dolev, S. Why future supercomputing requires optics. *Nat. Photon.* **4**, 261–263 (2010).
- Zhang, C., Zhang, S., Peters, J. D. & Bowers, J. E.  $8 \times 8 \times 40$  Gbps fully integrated silicon photonic network on chip. *Optica* **3**, 785–786 (2016).
- Shen, Y. W. et al. Silicon photonics for extreme scale systems. *J. Lightwave Technol.* **37**, 245–259 (2019).
- Wade, M. et al. In *2018 European Conference on Optical Communication (ECOC)*. 1–3 (IEEE, 2018).
- Wuttig, M., Bhaskaran, H. & Taubner, T. Phase-change materials for non-volatile photonic applications. *Nat. Photon.* **11**, 465–476 (2017).
- Yang, Z. & Ramanathan, S. Breakthroughs in photonics 2014: phase change materials for photonics. *IEEE Photon. J.* **7**, 1–5 (2015).
- Zhang, W., Mazzarello, R., Wuttig, M. & Ma, E. Designing crystallization in phase-change materials for universal memory and neuro-inspired computing. *Nat. Rev. Mater.* **4**, 150–168 (2019).
- Briggs, R. M., Pryce, I. M. & Atwater, H. A. Compact silicon photonic waveguide modulator based on the vanadium dioxide metal-insulator phase transition. *Opt. Express* **18**, 11192–11201 (2010).
- Wang, Q. et al. Optically reconfigurable metasurfaces and photonic devices based on phase change materials. *Nat. Photon.* **10**, 60–U75 (2016).
- Chu, C. H. et al. Active dielectric metasurface based on phase-change medium. *Laser Photon. Rev.* **10**, 986–994 (2016).
- Yin, X. et al. Beam switching and bifocal zoom lensing using active plasmonic metasurfaces. *Light. Sci. Appl.* **6**, e17016 (2017).
- Wu, C. et al. Low-loss integrated photonic switch using subwavelength patterned phase change material. *ACS Photon.* **6**, 87–92 (2018).
- Cheng, Z. et al. Device-level photonic memories and logic applications using phase-change materials. *Adv. Mater.* **30**, e1802435 (2018).
- Xu, P., Zheng, J., Doyle, J. K. & Majumdar, A. Low-loss and broadband nonvolatile phase-change directional coupler switches. *ACS Photon.* **6**, 553–557 (2019).
- Zhang, Y. et al. Broadband transparent optical phase change materials for high-performance nonvolatile photonics. *Nat. Commun.* **10**, 4279 (2019).
- de Galarreta, C. R. et al. Nonvolatile reconfigurable phase-change metadevices for beam steering in the near infrared. *Adv. Funct. Mater.* **28**, 1704993 (2018).
- Stegmaier, M., Ríos, C., Bhaskaran, H., Wright, C. D. & Pernice, W. H. P. Nonvolatile all-optical  $1 \times 2$  switch for chipscale photonic networks. *Adv. Opt. Mater.* **5**, 1600346 (2017).
- Zhang, Q. et al. Broadband nonvolatile photonic switching based on optical phase change materials: beyond the classical figure-of-merit. *Opt. Lett.* **43**, 94 (2018).
- Li, X. et al. Fast and reliable storage using a 5-bit, non-volatile photonic memory cell. *Optica* **6**, 1–6 (2019).
- Martins, T. et al. Fiber-integrated phase-change reconfigurable optical attenuator. *Appl. Photon.* **4**, 111301 (2019).
- Zheng, J. et al. Nonvolatile electrically reconfigurable integrated photonic switch enabled by a silicon PIN diode heater. *Adv. Mater.* **32**, e2001218 (2020).
- George, J. K. et al. Neuromorphic photonics with electro-absorption modulators. *Opt. Express* **27**, 5181–5191 (2019).
- Tait, A. N. et al. Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 064043 (2019).
- Hamerly, R., Bernstein, L., Sludds, A., Soljacic, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
- Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
- Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493**, 195–199 (2013).
- Ribeiro, A., Ruocco, A., Vanacker, L. & Bogaerts, W. Demonstration of a  $4 \times 4$ -port universal linear circuit. *Optica* **3**, 1348–1357 (2016).
- Bocker, R. P. Matrix multiplication using incoherent optical techniques. *Appl. Opt.* **13**, 1670–1676 (1974).
- Ríos, C. et al. In-memory computing on a photonic platform. *Sci. Adv.* **5**, eaau5759 (2019).
- Chakraborty, I., Saha, G. & Roy, K. Photonic in-memory computing primitive for spiking neural networks using phase-change materials. *Phys. Rev. Appl.* **11**, 014063 (2019).
- Caulfield, H. J., Kinsler, J. & Rogers, S. K. Optical neural networks. *Proc. IEEE* **77**, 1573–1583 (1989).
- Feldmann, J. et al. Parallel convolution processing using an integrated photonic tensor core. arXiv preprint arXiv:2002.00281 (2020).
- Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
- Li, Z. et al. Controlling propagation and coupling of waveguide modes using phase-gradient metasurfaces. *Nat. Nanotechnol.* **12**, 675–683 (2017).
- Park, J.-W. et al. Optical properties of pseudobinary GeTe, Ge<sub>2</sub>Sb<sub>2</sub>Te<sub>5</sub>, GeSb<sub>2</sub>Te<sub>4</sub>, GeSb<sub>4</sub>Te<sub>7</sub>, and Sb<sub>2</sub>Te<sub>3</sub> from ellipsometry and density functional theory. *Phys. Rev. B* **80**, 115209 (2009).
- Liu, Y., Aziz, M. M., Shalini, A., Wright, C. D. & Hicken, R. J. Crystallization of Ge<sub>2</sub>Sb<sub>2</sub>Te<sub>5</sub> films by amplified femtosecond optical pulses. *J. Appl. Phys.* **112**, 123526 (2012).
- Farmakidis, N. et al. Plasmonic nanogap enhanced phase-change devices with dual electrical-optical functionality. *Sci. Adv.* **5**, eaaw2687 (2019).
- Zhang, H. et al. Miniature multilevel optical memristive switch using phase change material. *ACS Photon.* **6**, 2205–2212 (2019).
- Rodríguez-Hernández, G., Hosseini, P., Ríos, C., Wright, C. D. & Bhaskaran, H. Mixed-mode electro-optical operation of Ge<sub>2</sub>Sb<sub>2</sub>Te<sub>5</sub> nanoscale crossbar devices. *Adv. Electron. Mater.* **3**, 1700079 (2017).
- Ríos, C. et al. Integrated all-photonic non-volatile multi-level memory. *Nat. Photon.* **9**, 725–732 (2015).
- Giannopoulos, I. et al. In *2018 IEEE International Electron Devices Meeting (IEDM)*. 27.27.21–27.27.24 (IEEE, 2018).
- Le Gallo, M. et al. Mixed-precision in-memory computing. *Nat. Electron.* **1**, 246–253 (2018).
- Nahmias, M. A. et al. Photonic multiply-accumulate operations for neural networks. *IEEE J. Sel. Top. Quant.* **26**, 1–18 (2020).
- Convolution Neural Network - simple code - simple to use (MATLAB Central File Exchange, 2020).
- Xiong, C. et al. Monolithic 56 Gb/s silicon photonic pulse-amplitude modulation transmitter. *Optica* **3**, 1060–1065 (2016).
- Moazeni, S. et al. 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**, 3503–3516 (2017).
- Sawchuk, A. A. & Jenkins, B. K. In *Optical Computing*. 143–153 (International Society for Optics and Photonics 1986).
- Joshi, A. et al. In *2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*. 124–133 (IEEE, 2009).
- Khope, A. S. P. et al. Multi-wavelength selective crossbar switch. *Opt. Express* **27**, 5203–5216 (2019).
- Ohno, S., Toprasertpong, K., Takagi, S. & Takenaka, M. Si microring resonator crossbar arrays for deep learning accelerator. *Jpn J. Appl. Phys.* **59**, SGGE04 (2020).
- Nahmias, M. A. et al. Photonic multiply-accumulate operations for neural networks. *IEEE J. Sel. Top. Quant. Electron.* **26**, 1–18 (2019).
- Han, S., Seok, T. J., Quack, N., Yoo, B.-W. & Wu, M. C. Large-scale silicon photonic switches with movable directional couplers. *Optica* **2**, 370–375 (2015).
- Tait, A. N., Chang, J., Shastri, B. J., Nahmias, M. A. & Prucnal, P. R. Demonstration of WDM weighted addition for principal component analysis. *Opt. Express* **23**, 12758 (2015).
- Tait, A. N., Nahmias, M. A., Shastri, B. J. & Prucnal, P. R. Broadcast and weight: an integrated network for scalable photonic spike processing. *J. Lightwave Technol.* **32**, 4029–4041 (2014).
- Silver, D. et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362**, 1140–1144 (2018).
- Jouppi, N. P. et al. In *Proceedings of the 44th Annual International Symposium on Computer Architecture - ISCA '17 1-12* (ACM Press, New York, New York, USA, 2017).
- Li, X. et al. Experimental investigation of silicon and silicon nitride platforms for phase-change photonic in-memory computing. *Optica* **7**, 218–225 (2020).
- Gayen, D. K., Chattopadhyay, T., Pal, R. K. & Roy, J. N. All-optical Multiplication with the help of Semiconductor Optical Amplifier—assisted Sagnac Switch. *J. Comput. Electron.* **9**, 57–67 (2010).
- Atabaki, A. H. et al. Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip. *Nature* **556**, 349–354 (2018).
- Bangari, V. et al. Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs). *IEEE J. Sel. Top. Quant.* **26**, 1–13 (2020).

## Acknowledgements

We acknowledge the funding support provided by the ONR MURI (Award No. N00014-17-1-2661). Part of this work was conducted at the Washington Nanofabrication Facility/Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure (NNCI) site at the University of Washington, which is supported in part by funds from the National Science Foundation (awards NNCI-1542101, 1337840, and 0335765), the National Institutes of Health, the Molecular Engineering & Sciences Institute, the Clean Energy Institute, the Washington Research Foundation, the M. J. Murdock Charitable Trust, Altatech, ClassOne Technology, GCE Market, and Google and SPTS.

## Author contributions

C.W. and M.L. conceived the research. C.W. fabricated the devices, performed the measurements and analysed the data. H.Y. and I.T. deposited the  $\text{Ge}_2\text{Sb}_2\text{Te}_5$  thin films and characterized the optical properties. S.L. and R.P. assisted the fabrication and characterization. C.W., I.T., and M.L. co-wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-20365-z>.

**Correspondence** and requests for materials should be addressed to M.L.

**Peer review information** *Nature Communications* thanks Wolfram Pernice and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission 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 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021