ARTICLE Open Access

SUANPAN: scalable photonic linear vector machine

Ziyue Yang[1], Chen Li[1], Yuqia Ran[2], Yongzhuo Li[1]✉, Xue Feng[1]✉, Kaiyu Cui[1], Fang Liu[1], Hao Sun[1], Wei Zhang[1], Yu Ye[2], Fei Qiao[1], Jiaxing Wang[3], Cun-Zheng Ning[1,4], Connie J. Chang-Hasnain[3] and Yidong Huang[1]✉

Abstract

Photonics is promising to handle extensive vector multiplications in artificial intelligence (AI) techniques due to natural bosonic parallelism and high-speed information transmission. However, the dimensionality of current photonic linear operation is limited and tough to improve due to the complex beam interaction for implementing optical matrix operation and digital-analog conversions. Here, we propose a programmable and reconfigurable photonic linear vector machine with extreme scalability formed by a series of emitter-detector pairs as the independent basic computing units. The elemental values of two high-dimensional vectors are prepared on emitter-detector pairs by bit encoding and analog detecting method without requiring large-scale analog-to-digital converter or digital-to-analog converter arrays. Since there is no interaction among light beams inside, extreme scalability could be achieved by simply multiplicating the independent emitter-detector pair. The proposed architecture is inspired by the traditional Chinese Suanpan or abacus, and thus is denoted as photonic SUANPAN. Experimentally, the computing fidelities for vector inner products could achieve >98% in our implementation with an 8×8 vertical cavity surface emission laser (VCSEL) array and an 8×8 $MoTe_2$ two-dimensional material photodetector array. Furthermore, such implementation is applied on two typical AI tasks as 1024-dimensional optimization problem is successfully solved and competitive classification accuracy of 88% is achieved for handwritten digit dataset. We believe that the photonic SUANPAN could serve as a fundamental linear vector machine and enhance various future AI applications.

Introduction

Artificial intelligence (AI) is currently an active topic in both scientific research and commercial applications as well as daily life[1,2]. The linear operations of high-dimensional vectors are fundamental and dominant in both the artificial neural networks[3–5] (ANN) and optimization problem solvers, *such as* the Ising machine[6–8]. As the complexity of problems increases, the dimensionality of the processed vector grows rapidly, resulting in a huge computational burden. It is known that vector operations can be readily accelerated by photons due to the natural parallelism of bosons[9]. In the past decades, various photonic computing architectures have been demonstrated to perform vector matrix multiplication in the optical domain, i.e., Stanford structure[10–12], Reck scheme[13–16], deep diffraction architecture[17–20], micro-ring resonator (MRR) array[21–26], etc. All these architectures perform vector matrix multiplications based on the interaction between light beams, which refers to coherent or incoherent superposition between different light beams through beam splitting, beam combining, diffracting, scattering, etc. However, as the optical matrix transformation is adopted, the basic units in the computing architecture, i.e., liquid crystal cells, beam splitters, meta-atoms, etc., would be tightly interconnected or highly coupled with each other due to the interaction. Thus, high-dimensional optical vector-matrix operations cannot be achieved by simply multiplicating these basic units, which significantly limits the scalability of the architecture. Here, instead of optical matrix operations, we propose the SUANPAN architecture for the optical

Correspondence: Yongzhuo Li (liyongzhuo@tsinghua.edu.cn) or
Xue Feng (x-feng@tsinghua.edu.cn) or
Yidong Huang (yidonghuang@tsinghua.edu.cn)
[1]Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China
[2]State Key Laboratory for Mesoscopic Physics and Frontiers Science Center for Nano-Optoelectronics, School of Physics, Peking University, 100871 Beijing, China
Full list of author information is available at the end of the article
These authors contributed equally: Ziyue Yang, Chen Li.

© The Author(s) 2026
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 licence, and indicate if changes were made. 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/4.0/.

inner product of two vectors. Just like the transistors in an integrated circuit, the independent basic computing unit in our scheme contains only one emitter-detector pair and could be scaled up to form a photonic computing chip. The elemental values of two vectors are encoded on the output intensity of the light-emitters and the photoresponsivity of the photodetectors (PDs), respectively. Thus, the photocurrent of the PD would be proportional to the multiplication of the light intensity and photoresponsivity, and the final result of the inner product can be obtained by the summation of all the photocurrents. Since there is no interaction among the propagating light beams of all emitter-detector pairs and only the output currents of all PDs are connected, our scheme is scalable by increasing the number of emitter-detector pairs with no additional loss or error, as well as flexibly reconfigurable and programmable for different computational tasks.

As a proof of principle, the SUANPAN architecture is implemented by utilizing an 8×8 vertical cavity surface emission laser (VCSEL) array and an 8×8 $MoTe_2$ two-dimensional (2D) material PD array. In the experiment, the calculation fidelity of the random vector inner product can be as high as >98% for various bit precisions (2-bit, 4-bit, and 8-bit), and >95% for various vector dimensionalities (@4-bit precision). Furthermore, such implementation has been successfully reconfigured to perform two typical AI tasks, the Ising machine and the ANN. A randomly generated 1024-dimensional Ising problem is successfully solved, which is the highest dimensionality of optical Ising machine with heuristic algorithm. Meanwhile, a competitive classification accuracy of 88% is achieved for ANN on the MNIST handwritten digit dataset. We believe that our proposed photonic SUANPAN is capable to serve as a fundamental linear vector machine and is potential to enhance the computing power for future various AI applications.

SUANPAN architecture

The proposed SUANPAN architecture consists of a light-emitter array and a PD array, as well as some necessary electronic hardware, as schematically shown in Fig.1a. To perform a vector inner product, both multiply and accumulate operations are required. Firstly, for multiply operation, each PD is well aligned with a corresponding light-emitter to form an emitter-detector pair, therefore, the photocurrent of the PD would be proportional to the multiplication of the light intensity and photoresponsivity due to the linear optical response[27]. Then, for add operation, all of the PDs are connected so that the output current would be the sum of all PDs due to Kirchhoff's law. In this way, the multiply-accumulate operation is naturally performed through the emission and detection. Then the key issue is how to encode the vectors on the emitter-detector pairs. A natural way is to directly encode on light intensity and photoresponsivity. However, it would be required that each PD and each emitter should be equipped with a high-precision and high-speed digital-to-analog converter (DAC), which would introduce large power and area overhead[28] as well as significant latency. Here, we have proposed the Bit Encoding and Analog Detecting paradigm to avoid DAC, thus, one emitter-detector pair is denoted as BEAD.

For deep insight, one BEAD is first considered. As shown in Fig. 2a, the multiplier a is encoded on the intensity of the light-emitter by controlling the duty ratio of driving current, which is done by a digital counter according to the clock cycles without DAC (the details of encoding shown in Supplementary Note 1 and Fig. S1). The bit precision depends on the time-slot numbers within the period. The multiplier b is encoded on the on-off state of the BEAD, by turning on (green arrow) or off (gray arrow) the light-emitter, respectively. Hence, there are two states to encode b, $b = 0$ or $b = 1$, known as 1-bit quantization, and the photocurrent would be proportional to $a \times b$, for $b = 0, 1$. For more bit quantization of b, more BEADs are employed to form a set. Considering 2-bit quantization, two BEADs should be employed in one set to obtain four combinations of on-off states. Thus, the two bits in the binary representation of b would be corresponding to these two BEADs as shown in Fig. 2b. For example, if $b = 2$, the binary representation would be $b = 10$, which means the first BEAD is at off-state and the second one is at on-state. This operation is quite similar to the Chinese traditional Suanpan, which represents numbers according to the position of beads and carry out mathematic operations by moving the beads up and down. Thus, our scheme is named as photonic SUANPAN. Moreover, different bits represent different weights in binary representation, which can be achieved by setting the photoresponsivity of two PDs as 2^0–2^1. To properly manipulate the photoresponsivity, 2D material photoconductive detectors are fabricated. By combining the photocurrents of two PDs, the output result is $a \times b$ as shown in Fig. 2b. In this way, M-bit quantization of b can be achieved with a set of M BEADs as shown in Fig. 2c, so that the SUANPAN can encode the range of b from 0 to $2^M - 1$ and achieve the multiplication of a and b.

Since a is digitally encoded to the duty ratio of light emission and b is digitally encoded to the on-off states of the BEADs in one set, while the output result is the total analog photocurrent from all the PDs, the emitter-detector pair is actually operated as Bit Encoding and Analog Detecting. Also, the numbers of a and b are encoded in time and space domain respectively, therefore, the SUANPAN architecture can perform the multiplication of any desired bit precision theoretically.

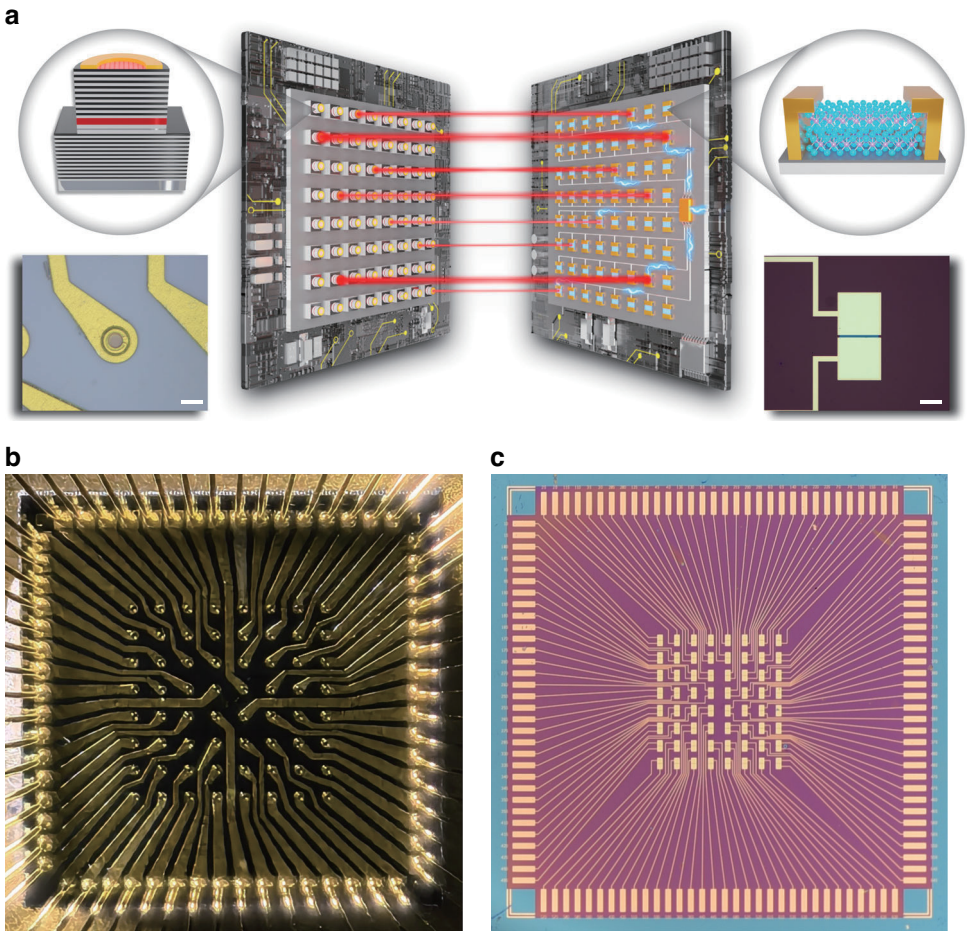

Fig. 1 Architecture of the SUANPAN. a The schematic diagram of the SUANPAN architecture, consisting of a light-emitter array, a PD array, and some necessary electronic hardware. Left insets show the schematic and microscope photograph of a single VCSEL. Scale bar is 20 μm. The right insets show the schematic and microscope photograph of a single $MoTe_2$ PD. Scale bar is 100 μm. **b** The optical image of the VCSEL array. **c** The optical image of the $MoTe_2$ PD array

Figure 2d shows an example of the time-space encoding, where a is between 0 and 100 by controlling the duty ratio as 0–100%, and b is 4-bit quantization by employing a set of 4 BEADs.

It should be mentioned that the negative numbers can also be handled by applying reversed bias voltage of the PD. Considering both positive and negative numbers, $2 \times M$ BEADs are required for each set as shown in Supplementary Note 2 and Fig. S2. Furthermore, the complex vector inner product can also be handled as shown in Supplementary Note 3 and Fig. S3. Thus, the SUANPAN architecture could achieve both reconfigurable and programmable ability, since the number of BEADs in each set can be reconfigured according to the bit precision, while the exact encoding of each BEAD can be flexibly programmed according to the elements within the vectors. At the output, the photocurrents of all PDs are connected, so that only one ADC is required to transform the total photocurrent into a digital signal. Also, since only 1-bit information would be encoded on a single BEAD, the properly settled but fixed bias voltage would be applied. Thus, there is no requirement for DAC in the SUANPAN. Last but not least, since there is no interaction among the propagating light beams of all BEADs, the SUANPAN architecture is scalable by increasing the number of independent BEADs with no additional loss or error. On one hand, the number of utilized BEADs can be increased by integrating more light-emitters and PDs on one chip. On the other hand, distributed computing can also be achieved by simply connecting multiple chips together to scale up the computing power more. Therefore, the SUANPAN is a programmable, reconfigurable, and scalable architecture, which can serve as a general vector inner product accelerator for the existing electronic computing system.

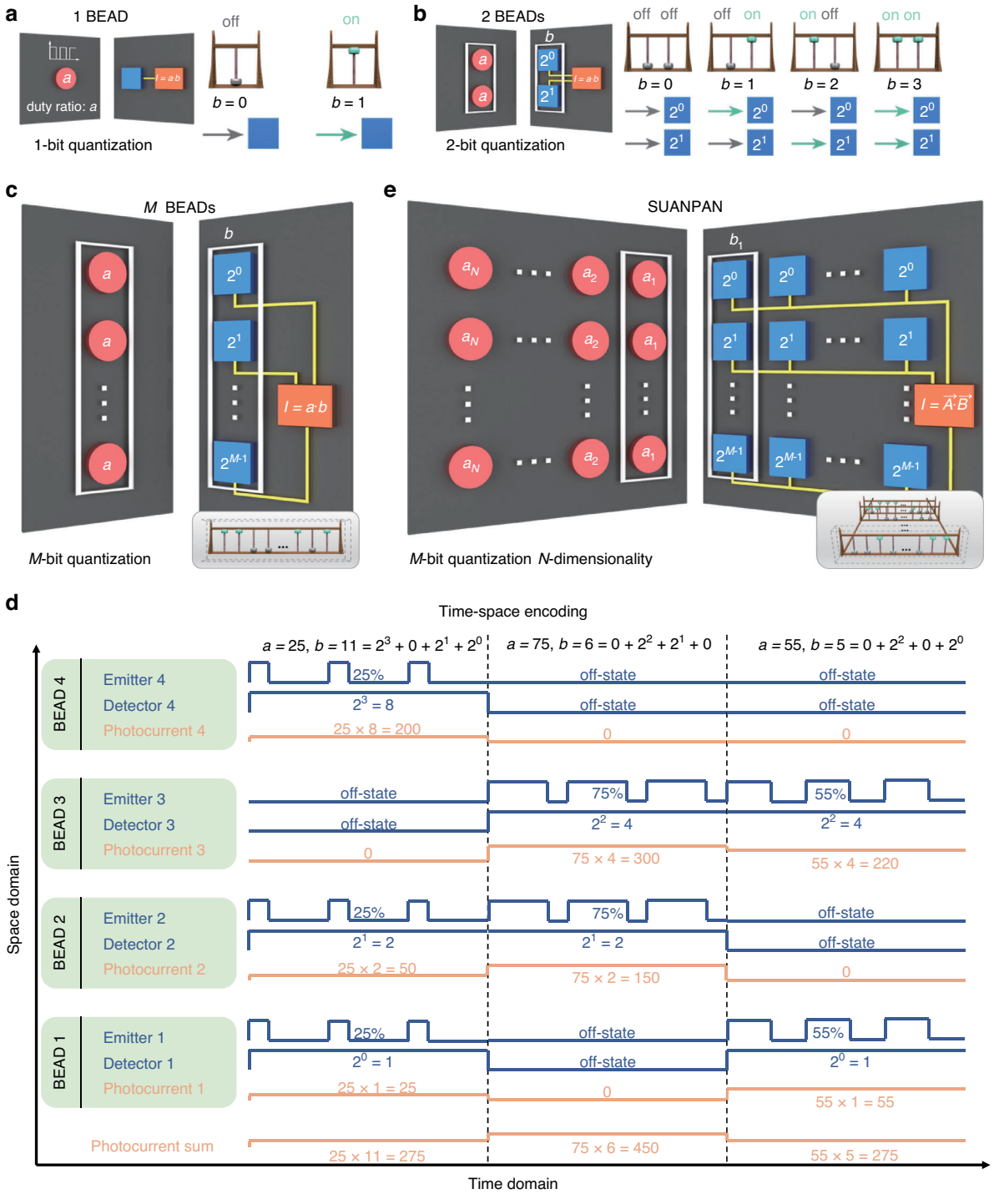

Fig. 2 Operation principle of the SUANPAN. a The mechanism of performing $a \times b$, where b is 1-bit quantization with 1 BEAD. Right inset: the on-off state of 1 BEAD in the SUANPAN comparing with 1 bead in Chinese Suanpan. **b** The mechanism of performing $a \times b$, where b is 2-bit quantization with 2 BEADs. Right inset: the on-off states of 2 BEADs in the SUANPAN comparing with 2 beads in the Chinese Suanpan. **c** The mechanism of performing $a \times b$, where b is M-bit quantization with M BEADs. The white box represents a set of M BEADs. Inset: M beads in Chinese Suanpan as a comparison. **d** An example for the time-space encoding, where $a = 0, 1, \ldots, 100$ by controlling the duty ratio as 0–100%, and b is 4-bit quantization by employing 4 BEADs. Since all these 4 detectors are connected, the total photocurrent is proportional to $a \times b$. **e** The mechanism of performing vector inner product $A{\cdot}B$ with N sets of M BEADs. Inset: N sets of Chinese Suanpan as a comparison

Results

To implement the prototype of the SUANPAN, a pair of VCSEL and $MoTe_2$ PD is employed to form the BEAD. The schematic diagram and microscope photographs of a single VCSEL and $MoTe_2$ PD are shown in the insets of Fig. 1a. As a light-emitter, VCSEL can readily achieve high-speed modulation as well as a large-scale array. Recently, researchers have already demonstrated a neural network based on VCSEL[29]. While, in such architecture, each VCSEL requires injection to achieve a stable phase lock. In comparison, for the SUANPAN architecture, all VCSELs are independent. Thus, the phase locking, as well as other additional operations are not required. For PD, 2D material is utilized for three reasons: (1) The photoresponsivity of 2D material PD can be flexibly controlled by the bias voltage. (2) The high carrier mobility in 2D material[30] can support high-speed detection, which is an important issue for high-speed computing. (3) 2D material can be heterogeneously integrated with other material platform[31], therefore, 2D material PD is potentially integrated with a light-emitter in the future. Specifically, according to the wavelength of VCSEL (850 nm), $MoTe_2$ is utilized as the PD material. Thus, we have fabricated both the VCSEL and $MoTe_2$ PD array chip with 8 × 8 components as shown in Fig. 1b and Fig. 1c, respectively. The fabrication process, experimental setup, and performance of VCSEL and PD are shown in Figs. S4–S12 and discussed in Methods. Actually, the BEAD can be achieved with various emitter and detector combinations. Both laser and light emitter diode (LED) could serve as the emitter, while different 2D materials, i.e., graphene[32], MoS_2[33,34], WSe_2[35–37], etc. could be applied for PD according to the proper operation wavelength.

To verify the functionality of the SUANPAN, random vector inner products are firstly performed with bit precision of 2-bit, 4-bit, and 8-bit. For a 2-bit precision signed vector, 4 BEADs are required in each set so that a 16-dimensional vector inner product can be done at one time. Similarly, for 4-bit and 8-bit quantization, the corresponding dimensionality would be 8 and 4, respectively. To achieve higher-dimensional vector inner product, time-division multiplexing can also be employed. For each bit precision, the configuration of the SUANPAN would be properly settled and the corresponding bias voltage of each PD is shown in Fig. 3a–c, respectively. Also, 1000 rounds of signed vector inner products are randomly generated and performed by the SUANPAN. To evaluate the accuracy, the normalized results of 1000 rounds for each bit precision calculated by the SUANPAN and computer are shown in Fig. 3d–f, respectively. The achieved fidelities are all higher than 98% (details provided in Supplementary Note 5), which indicates that the SUANPAN can perform accurate calculation.

Specifically, for 4-bit precision, the experimental photocurrent for $a \times b$ is shown in Fig. 3g, where a remains unchanged and b takes 0, ±1, …, ±15. The recorded on-off states of the utilized 8 BEADs are also shown in the bottom inset of Fig. 3g, which is corresponding to the first column of Fig. 3b. Moreover, the fidelities of 4-bit precision with various dimensionalities are shown in Fig. 3h. As the dimensionality increases, the computational fidelity remains above 95%. Due to the high fidelity in executing random signed vector inner products, the SUANPAN architecture can be flexibly utilized to further demonstrate more specific computing tasks. Here, two typical AI tasks are considered, the Ising problem[38] and ANN, in which the vector multiplication is the core computing operation as shown in Fig. 4a.

An N-dimensional Ising problem is defined by a symmetric interaction matrix J ($N \times N$ dimensionality with diagonal elements of zero), and the Hamiltonian of Ising problem is defined as follows:

$$H = S^{\mathrm{T}} J S \tag{1}$$

Solving Ising problem is to find the specific vector S that minimizes the Hamiltonian, which is denoted as the ground state. Here, the simulated annealing (SA) algorithm[39] is employed, which searches for the ground state through multiple iterations. In each iteration of SA, the variation of Hamiltonian ΔH is calculated, which can be transformed into an N-dimensional vector inner product and can be readily performed by the SUANPAN (details in Supplementary Note 6). According to the previous reports[40], a programmable photonic Ising machine with heuristic algorithms has successfully solved the highest dimensionality of 30-dimensional arbitrarily connected Ising problem (detailed discussion and comparison can be found in our previous work[40]). For convenient comparison, a randomly generated 30-dimensional Ising problem is solved experimentally by the SUANPAN. The solving process is repeated with 100 rounds, and the 100 annealing curves are shown in Fig. 4b. It can be seen that 99 curves eventually converged to the ground state (dashed line shown in Fig. 4b), therefore, an accuracy of 99% is achieved by the SUANPAN, which is much higher than the existing 30-dimensional Ising machine based on SA[40]. Further, a randomly generated 1024-dimensional Ising problem is considered. As common practices, an approximate solution with 87.8% of the ground state is set as a criterion for successful solution[41–43], as dashed line shown in Fig. 4c. Such 1024-dimensional Ising problem is successfully solved by the SUANPAN as the annealing curve fall below the criterion line after ~4000 iterations. The high convergence rate and high dimensionality for solving various Ising problems can validate the

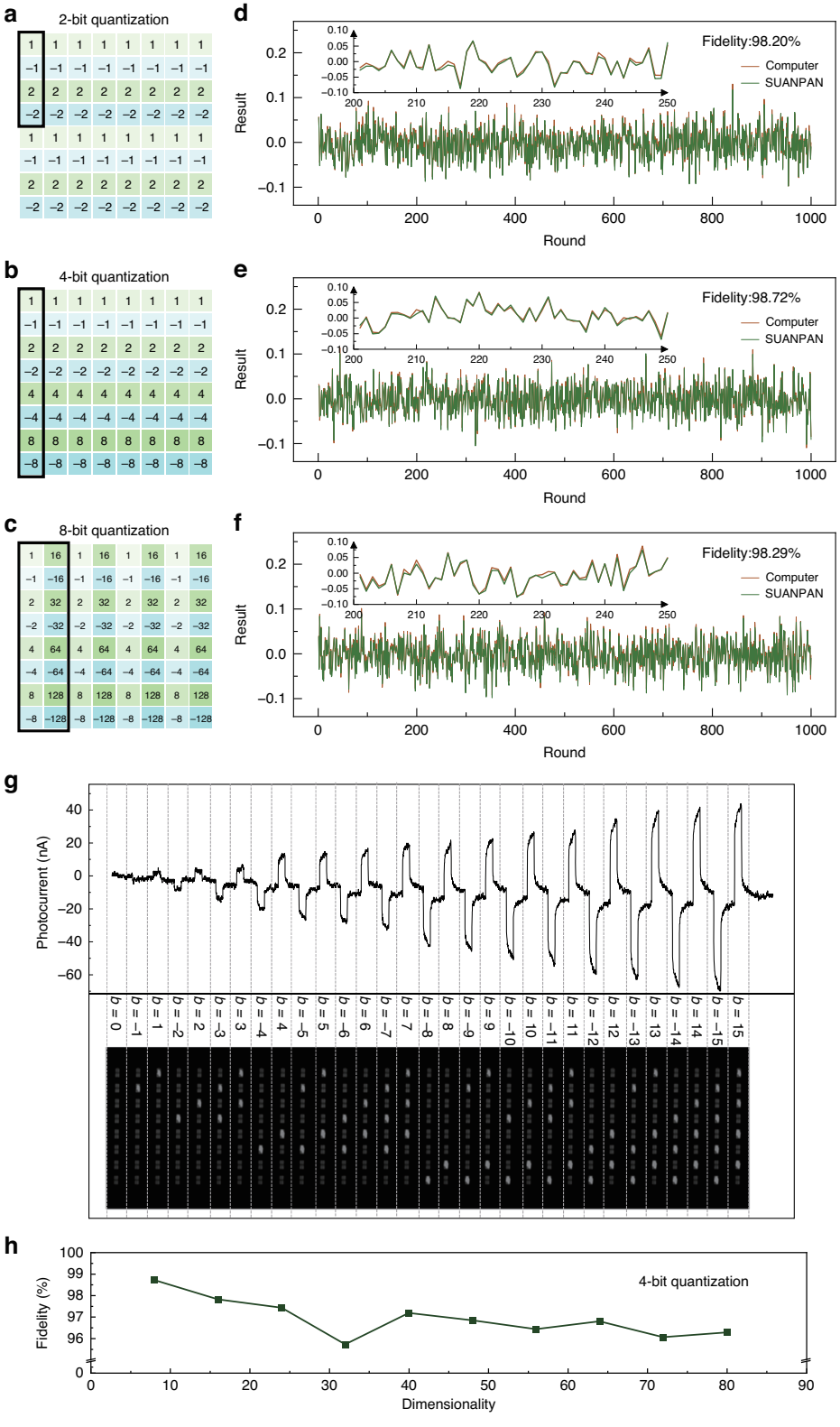

Fig. 3 Random vector inner product. a–c The configuration of the SUANPAN for 2-bit, 4-bit and 8-bit quantization, respectively. **d–f** The normalized results calculated by the SUANPAN and computer for 2-bit, 4-bit and 8-bit quantization, respectively. Insets: the enlarged view of round 201–250. **g** The experimental results of $a \times b$, where a remains unchanged and b takes each value in 4-bit quantization. **h** The experimental fidelity for 4-bit quantization with various dimensionalities

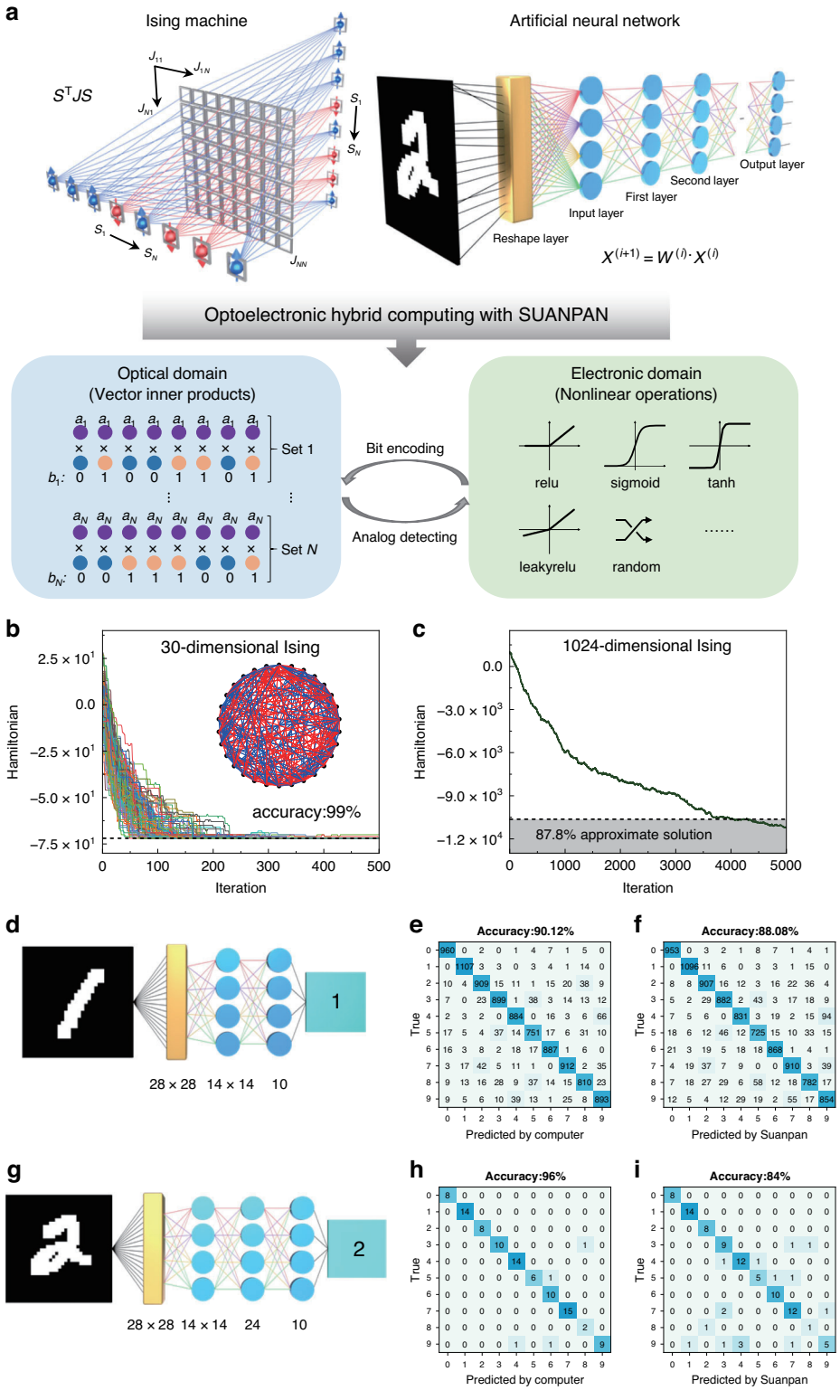

Fig. 4 AI applications. a Various AI applications can be performed on the SUANPAN through optoelectronic hybrid computing. **b** 100 experimental annealing curves of a random 30-dimensional Ising problem. Dashed line: ground state. Inset: the random 30-dimensional Ising model (red: $J_{ij} = 1$, blue: $J_{ij} = -1$). **c** An experimental annealing curve of a random 1024-dimensional Ising problem. Dashed line: 87.8% approximate solution. **d** The pre-trained single-layer ANN for MNIST dataset. **e**, **f** The accuracy and confusion matrix of the single-layer ANN performed by computer and the SUANPAN, respectively. **g** The pre-trained double-layer ANN for MNIST dataset. **h**, **i** The accuracy and confusion matrix of the double-layer ANN performed by computer and the SUANPAN, respectively

programmability, reconfigurability and computational stability of the SUANPAN architecture. Recently, an on-chip Ising machine[44] is demonstrated with both linear and nonlinear operations in optical domain. Of course, it is beyond the scope of this work. But it is an interesting topic to combine the SUANPAN architecture with nonlinearity, and we are still undergoing it.

For ANN, various physical neural networks (PNNs), including optical neural networks (ONNs), have been applied to accelerate the calculation. In such PNNs, silico training is usually required to avoid errors caused by differences between simulation and practical devices. Unlike that, the SUANPAN can directly map a pre-trained ANN, in which the vector matrix multiplication can be considered as a set of vector inner products and performed in the optical domain, while the nonlinear activation function would be executed by an electronic processor. Therefore, through time-division multiplexing, the SUANPAN can execute ANNs of varied depth and number of nodes in theory. It should be mentioned that it depends on both reconfigurable and programmable abilities of the SUANPAN since the dimensionality cannot be extended with time-division multiplexing for a fixed computing architecture. Here, both single and double layer ANNs are performed as shown in Figs. 4d and g, respectively. MNIST handwritten digit dataset is utilized as dataset, and stochastic gradient descent[45] (SGD) is utilized as training method. The weights of the single-layer ANN and double-layer ANN are 4-bit precision and 6-bit precision according to simulations, respectively (the details are shown in Supplementary Note 7 and Fig. S13). For single-layer ANN, the confusion matrix of 10,000 pictures in the test dataset calculated by computer and the SUANPAN are shown in Fig. 4e, f, respectively. The approaching classification accuracies are 88.08% and 90.12% for the SUANPAN and computer, respectively. It can be seen that the experimental accuracy is 98% of the simulation accuracy, which is comparable with the previous work[29]. This result indicates that only a little deterioration is introduced by the SUANPAN. While for double-layer, only the first 100 pictures in MNIST test dataset are performed as a preliminary verification. The confusion matrix and accuracy calculated by computer and the SUANPAN are shown in Fig. 4h, i, respectively. It can be noticed that the classification accuracy calculated by the computer is much higher than one-layer, while that calculated by the SUANPAN is lower than one-layer. The reason might be the performance of $MoTe_2$ PD array has deteriorated after three months of testing (the details are shown in Supplementary Note 12 and Fig. S16). Anyway, we believe that the above results of ANNs can still validate the feasibility of the SUANPAN architecture.

Discussion

In this work, we have proposed and demonstrated the photonic SUANPAN architecture to perform the vector inner product operations. As a proof of principle, a SUANPAN with 64 pairs of VCSEL and $MoTe_2$ PD is implemented. According to the experimental results, the SUANPAN is capable of achieving high computing fidelities for randomly generated vector inner products, and can be applied on two typical AI tasks of the Ising machine and ANN. There are two main contributions in this work.

Firstly, for the SUANPAN architecture, it breaks through the traditional mindset of obtaining optical matrix transformations through the interaction of light beams. Instead, there is no interaction among those propagating light beams of all BEADs. Therefore, the SUANPAN can be decomposed into BEADs as independent computing units. The scalability, reconfigurability, and programmability of the SUANPAN architecture are only based on the multiplication, recombination and modulation of BEAD without any additional cost. Compared with optical matrix transformations through interaction between light beams, the SUANPAN possesses the following advantages: (1) With massive and industrial multiplication of BEADs, the SUANPAN can theoretically be infinitely scalable. (2) The SUANPAN can be flexibly reconfigured and programmed to perform various specific computing tasks. (3) Only correcting the intensity of light beam is required (details are provided in Supplementary Note 8), and there is no requirement to correct the phase term. (4) Even if one BEAD is broken during fabrication or operation, other BEADs would not be affected, and only the operating dimensionality would be decreased. Although 64 BEADs are operating well in our experimental demonstration, the anti-failure ability of the SUANPAN architecture would be of great significance for future large-scale computing since the yield of massive production cannot be always as 100%.

Secondly, the SUANPAN provides a promising solution for optoelectronic analog-digital hybrid computing. Large-scale DAC and ADC arrays are usually required in optoelectronic computing. However, with Bit Encoding and Analog Detecting paradigm, M-bit digital electronic signal is converted to analog within a set of M BEADs, while each BEAD only represents 1-bit information. Thus, no DAC is required. At the same time, only one ADC is required to convert the total photocurrent into electronic digital signal. Therefore, the Bit Encoding and Analog Detecting computing paradigm greatly reduces the heavy burden introduced by ADC and DAC. Actually, it is also an important issue for the scalability of the SUANPAN architecture.

Thirdly, we would provide a detailed analysis about the energy consumption. The energy consumption of the

SUANPAN would be approximately proportional to the number of bit precision, since each BEAD only encoding 1-bit information, and M BEADs are required for M-bit quantization. The energy consumption of a single BEAD consists of two parts: the energy consumption of VCSEL and that of the $MoTe_2$ PD. At 8-bit precision, the average energy consumption of a single VCSEL is ~2.5 mW, and the average energy consumption of a single PD is ~273.4 nW (the details are shown in Supplementary Note 11). Therefore, the total energy consumption of a BEAD is ~2.5 mW. It can be seen that the main energy consumption comes from the VCSEL. The reason might be due to the beam spreading during propagation, the channel of the PD only received a small part of the light beam, and the rest would be wasted (the radius of the light spot is ~200 μm, while the length of the channel is only 10 μm). If the PD and the light-emitter are integrated into a single chip in the future, the efficiency of the light power can be significantly improved.

Finally, we would provide a detailed analysis about the computing speed for both the current and predictable implementations of the photonic SUANPAN. Since all BEADs operate in parallel, the computing speed would not degrade as the bit precision increases. Considering one BEAD, the computing latency contains encoding time of the emitter (t_e), propagating time of light (t_p), and detecting time of the PD (t_d). The rise time and fall time of the VCSEL are 0.46 ns and 0.54 ns, respectively (as shown in Supplementary Note 9 and Fig. S14). In the current implementation, each VCSEL is operating at 100 MHz. Thus, the encoding time is $t_e = 1$ μs for multiplier $a = 0, 1, \ldots, 100$. The distance between VCSEL and PD is about 1.5 m, then the propagating time is about $t_p = 5$ ns. The rise time and fall time of the PD are 4.72 μs and 6.59 μs, respectively (as shown in Supplementary Note 10 and Fig. S15). Thus, the detecting time of the PD is $t_d = 6.59$ μs, which is the larger one between rise time and fall time. Then, the total computing latency would be $t_e + t_p + t_d = 7.59$ μs, and the computing speed for one BEAD would be 132 KOPS (operation per second). Since our implementation consists of 64 BEADs, considering 8-bit quantization of multiplier b, the computing speed of current SUANPAN would be 1.05 MOPS. Actually, the current implementation is only a prototype of the SUANPAN architecture, and there is still a lot of room to improve the performance. Obviously, both the light-emitter and the PD can be integrated into a single chip with the heterogeneous integration of 2D materials. Then, the imaging system in current setup is not required, and the propagating time would be greatly reduced. For example, if the distance between the emitter and PD is reduced to <1 mm, the propagating time would be $t_p < 3.3$ ps. Thus, the computing speed is mainly determined by two factors, the BEAD number and bandwidth. Similar to the development of integrated circuit, the BEAD number could be increased through continuously reducing the size of light-emitter and PD. Also, 3D multilayer stacking integration can be utilized to further expand the dimensionality. Meanwhile, due to the research on various high-speed nano-lasers[46–48] and nano-detectors[49,50], the bandwidth of a single BEAD, which is actually determined by the lower one between the emitter and PD, could be readily increased to several tens of gigahertz. As a concrete example, the computing speed could achieve >1 POPS/cm^2 (for 1-bit quantization) in a single chip for BEAD size <10 μm and BEAD bandwidth >1 GHz. It should be noticed that the aforementioned BEAD size and bandwidth are not very difficult to achieve. For example, the operation bandwidth of 50 GHz is achieved on $MoTe_2$ PD[51]. Furthermore, both the VCSEL and PD utilized in each BEAD are polarization-insensitive. If the polarization dimension encoding is also introduced into the SUANPAN architecture through properly manipulating the polarization state of both the emitter and the detector[52], the computing power could be boosted more. Therefore, we believe that the photonic SUANPAN architecture is very promising as an attractive and practicable linear vector machine in the visible future.

Materials and methods

Device fabrication of VCSEL

The 850 nm VCSEL epitaxy structure consists of around 35 pairs of AlGaAs bottom distributed Bragg reflector (DBR) and 25 pairs of AlGaAs top DBR. AlGaAs/InGaAs quantum well is used as the active region. 98% AlGaAs layer is used to form oxide aperture. After epitaxy, the wafer goes through P-metal deposition, inductively couple plasma trench and wet oxidation. The N-metal is deposited on the backside of the N-type substrate to form a common cathode. While the individual emitter in the array is connected to separate anode pads on the edge of the VCSEL array chip by electroplated traces (Fig. S11a).

Device fabrication of $MoTe_2$ PD

The PDs are fabricated on a SiO_2/Si substrate directly grown with a 10 nm 2H $MoTe_2$ layer (detailed fabrication process in refs.[53,54]). First, the patterns are defined by ultraviolet lithography and transferred to the $MoTe_2/SiO_2$ layer by reactive ion etching (SF_6 acts as the etching gas). Then, the Cr/Au electrodes (10 nm/50 nm) are fabricated using ultraviolet lithography, deposition and lift-off. The schematic diagram of the preparation process is shown in Fig. S4. To prevent degradation, the PDs are packaged with a 10 nm Al_2O_3 layer grown by atomic layer deposition. For subsequent testing, the $MoTe_2$ PDs are connected to a self-designed printed circuit board using wire-bonding technology (Fig. S11b).

Experimental setup

Schematics of the experimental setup are illustrated in Figs.1a and S12. Light from the VCSEL array with a wavelength of 850 nm is focused by a zoom lens onto the $MoTe_2$ PD array. The 8 × 8 VCSEL array and the PD array are aligned by the illumination optical path. Electrical and optoelectronic measurements of the fabricated $MoTe_2$ PDs are carried out with a semiconductor parameter analyzer (PDA FS380) at room temperature in ambient conditions. The time-resolved photoresponse of the PD is measured by the semiconductor parameter analyzer, and the modulation of the laser (S1FC635PM, 635 nm, Thorlabs) is realized through a function waveform generator (DG4062, RIGOL), which creates square wave pulses.

Acknowledgements
Funding from the National Key Research and Development Program of China (2023YFB2806703), the National Natural Science Foundation of China (Grant Nos. U22A6004, 92365210, and 62175124) is greatly acknowledged. This work was also supported by Beijing National Research Center for Information Science and Technology (BNRist), Frontier Science Center for Quantum Information, Beijing Academy of Quantum Information Science, and Tsinghua University Initiative Scientific Research Program.

Author details
[1]Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China. [2]State Key Laboratory for Mesoscopic Physics and Frontiers Science Center for Nano-Optoelectronics, School of Physics, Peking University, 100871 Beijing, China. [3]Berxel Photonics Company Ltd, 518071 Shenzhen, China. [4]College of Integrated Circuits and Optoelectronic Chips, Shenzhen Technology University, 518118 Shenzhen, China

Author contributions
Z.Y., Y.L., and X.F. conceived the idea. Z.Y. and C.L. designed and performed the simulations, experiments and data analysis. Y.R. and Y.Y. contributed to the growth of $MoTe_2$ layer on the SiO_2/Si substrate. J.W. and C.J.C.-H. contributed to the fabrication of the VCSEL array. F.Q. assisted in building the electronic control system of the SUANPAN architecture. H.S., K.C., F.L., W.Z., and C.N. provided useful discussions and comments. Z.Y., C.L., Y.L., X.F., and Y.H. wrote the paper. All authors revised and approved the manuscript.

Data availability
The data that support the findings of this study are available within the paper and the Supplementary. Other relevant data are available from the corresponding author on reasonable request.

Conflict of interest
Cun-Zheng Ning is an Editor for the journal, and no other author has reported any competing interests.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41377-025-02059-7.

Received: 9 April 2025 Revised: 11 September 2025 Accepted: 16 September 2025
Published online: 01 January 2026

References

1. Cetinic, E. & She, J. Understanding and creating art with AI: review and outlook. *ACM Trans. Multimed. Comput. Commun. Appl.* **18**, 66 (2022).
2. Rajpurkar, P. et al. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
4. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
5. Fu, T. Z. et al. Optical neural networks: progress and challenges. *Light Sci. Appl.* **13**, 263 (2024).
6. Mohseni, N., McMahon, P. L. & Byrnes, T. Ising machines as hardware solvers of combinatorial optimization problems. *Nat. Rev. Phys.* **4**, 363–379 (2022).
7. Laydevant, J., Marković, D. & Grollier, J. Training an Ising machine with equilibrium propagation. *Nat. Commun.* **15**, 3671 (2024).
8. Nikhar, S. et al. All-to-all reconfigurability with sparse and higher-order Ising machines. *Nat. Commun.* **15**, 8977 (2024).
9. Zhou, H. L. et al. Photonic matrix multiplication lights up photonic accelerator and beyond. *Light Sci. Appl.* **11**, 30 (2022).
10. 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).
11. Spall, J. et al. Fully reconfigurable coherent optical vector-matrix multiplication. *Opt. Lett.* **45**, 5752–5755 (2020).
12. Wang, T. Y. et al. An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).
13. Reck, M. et al. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73**, 58–61 (1994).
14. Shen, Y. C. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
15. Roques-Carmes, C. et al. Heuristic recurrent algorithms for photonic Ising machines. *Nat. Commun.* **11**, 249 (2020).
16. Pai, S. et al. Experimentally realized in situ backpropagation for deep learning in photonic neural networks. *Science* **380**, 398–404 (2023).
17. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
18. Yan, T. et al. Fourier-space diffractive deep neural network. *Phys. Rev. Lett.* **123**, 023901 (2019).
19. Zhou, T. K. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photonics* **15**, 367–373 (2021).
20. Fu, T. Z. et al. Photonic machine learning with on-chip diffractive optics. *Nat. Commun.* **14**, 70 (2023).
21. Tait, A. N. et al. Broadcast and weight: an integrated network for scalable photonic spike processing. *J. Light. Technol.* **32**, 4029–4041 (2014).
22. Deng, Y. B. & Chu, D. P. Coherence properties of different light sources and their effect on the image sharpness and speckle of holographic displays. *Sci. Rep.* **7**, 5893 (2017).
23. Feldmann, J. et al. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
24. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
25. Bai, B. W. et al. Microcomb-based integrated photonic processing unit. *Nat. Commun.* **14**, 66 (2023).
26. Dong, B. W. et al. Partial coherence enhances parallelized photonic computing. *Nature* **632**, 55–62 (2024).
27. Huo, N. J. & Konstantatos, G. Recent progress and future prospects of 2D-based photodetectors. *Adv. Mater.* **30**, 1801164 (2018).
28. Kim, S. et al. Neuro-CIM: ADC-less neuromorphic computing-in-memory processor with operation gating/stopping and digital–analog networks. *IEEE J. Solid-State Circuits* **58**, 2931–2945 (2023).
29. Chen, Z. J. et al. Deep learning with coherent VCSEL neural networks. *Nat. Photonics* **17**, 723–730 (2023).
30. An, J. R. et al. Perspectives of 2D materials for optoelectronic integration. *Adv. Funct. Mater.* **32**, 2110119 (2021).
31. You, J. et al. Hybrid/integrated silicon photonics based on 2D materials in optical communication nanosystems. *Laser Photonics Rev.* **14**, 2000239 (2020).
32. Koepfli, S. M. et al. Metamaterial graphene photodetector with bandwidth exceeding 500 gigahertz. *Science* **380**, 1169–1174 (2023).
33. Lopez-Sanchez, O. et al. Ultrasensitive photodetectors based on monolayer MoS_2. *Nat. Nanotechnol.* **8**, 497–501 (2013).
34. Qi, P. F. et al. Remote lightening and ultrafast transition: intrinsic modulation of exciton spatiotemporal dynamics in monolayer MoS_2. *ACS Nano* **14**, 6897–6905 (2020).
35. Baugher, B. W. H. et al. Optoelectronic devices based on electrically tunable p–n diodes in a monolayer dichalcogenide. *Nat. Nanotechnol.* **9**, 262–267 (2014).

36. Qi, P. F. et al. Phonon scattering and exciton localization: molding exciton flux in two dimensional disorder energy landscape. *eLight* **1**, 6 (2021).
37. Qi, P. F. et al. Giant excitonic upconverted emission from two-dimensional semiconductor in doubly resonant plasmonic nanocavity. *Light Sci. Appl.* **11**, 176 (2022).
38. Lucas, A. Ising formulations of many NP problems. *Front. Phys.* **2**, 5 (2014).
39. van Laarhoven, P. J. M. & Aarts, E. H. L. *Simulated Annealing: Theory and Applications* (Springer, 1987).
40. Ouyang, J. Y. et al. On-demand photonic Ising machine with simplified Hamiltonian calculation by phase encoding and intensity detection. *Commun. Phys.* **7**, 168 (2024).
41. Yamamoto, Y. et al. Coherent Ising machines—optical neural networks operating at the quantum limit. *npj Quantum Inf.* **3**, 49 (2017).
42. Haribara, Y., Utsunomiya, S. & Yamamoto, Y. Computational principle and performance evaluation of coherent ising machine based on degenerate optical parametric oscillator network. *Entropy* **18**, 151 (2016).
43. Goemans, M. X. & Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM* **42**, 1115–1145 (1995).
44. Wu, B. et al. A monolithically integrated optical Ising machine. *Nat. Commun.* **16**, 4296 (2025).
45. Ketkar, N. *Deep Learning with Python: A Hands-on Introduction* (Berkeley: Apress Berkeley, 2017).
46. Jeong, K. Y. et al. Recent progress in nanolaser technology. *Adv. Mater.* **32**, 2001996 (2020).
47. Du, W. et al. Nanolasers based on 2D materials. *Laser Photonics Rev.* **14**, 2000271 (2020).
48. Zhang, Q. et al. Halide perovskite semiconductor lasers: materials, cavity design, and low threshold. *Nano Lett.* **21**, 1903–1914 (2021).
49. Long, M. S. et al. Progress, challenges, and opportunities for 2D material based photodetectors. *Adv. Electron. Mater.* **29**, 1803807 (2019).
50. Liu, C. Y. et al. Silicon/2D-material photodetectors: from near-infrared to mid-infrared. *Light Sci. Appl.* **10**, 123 (2021).
51. Flöry, N. et al. Waveguide-integrated van der Waals heterostructure photodetector at telecom wavelengths with high speed and high responsivity. *Nat. Nanotechnol.* **15**, 118–124 (2020).
52. Li, R. Z. et al. On-chip metasurface-mediated $MoTe_2$ photodetector with electrically tunable polarization-sensitivity. *Adv. Opt. Mater.* **13**, 2402668 (2025).
53. Xu, X. L. et al. Millimeter-scale single-crystalline semiconducting $MoTe_2$ via solid-to-solid phase transformation. *J. Am. Chem. Soc.* **141**, 2128–2134 (2019).
54. Pan, Y. et al. Heteroepitaxy of semiconducting 2H-$MoTe_2$ thin films on arbitrary surfaces for large-scale heterogeneous integration. *Nat. Synth.* **1**, 701–708 (2022).