Supporting Information for

High-throughput optical neuromorphic graphic processing at millions of images per second

Yibo Dong[1,2], Yuchi Bai[1,2], Qiming Zhang[1,2,*], Haitao Luan[1,2,*], Min Gu[1,2,*]

[1]School of Artificial Intelligence Science and Technology, University of Shanghai for Science and Technology, Shanghai, 200093, China

[2]Institute of Photonic Chips, University of Shanghai for Science and Technology, Shanghai, 200093, China

*Email: qimingzhang@usst.edu.cn (Q. M. Z.); haitaoluan@usst.edu.cn (H. T. L.); gumin@usst.edu.cn (M. G.).

Supplementary Text

Supplementary Note 1: Calculation of the near-field pattern of the VCSEL based on the fiber model.

The structure of the oxide-confined VCSEL features a cylindrical design. From top to bottom, it can be divided into p-distributed Bragg reflector (DBR), oxidized layer, active region, n-DBR, n-GaAs substrate, respectively. After the oxidation process, the cladding formed by the oxidized region has a lower refractive index, while the core at the center maintains a higher refractive index, creating a cylindrical waveguide. In the oxidized region, the refractive index is around 1.65, significantly lower than the unoxidized region, which is about 2.97. This refractive index difference is confined to the oxidized layer, so the overall effective refractive index difference between the cladding and the core is approximately 0.02. As a result, the near-field pattern of the VCSEL can be calculated based on the fiber model with weak-guidance approximation[S1]. We should note that due to minor variations caused by manufacturing inconsistencies, different VCSELs exhibit slight differences in their near-field patterns. To simplify the model, we treat each VCSEL as an ideal single-mode light source. In the experiment, fundamental mode output is achieved by controlling the current injection.

In the fiber mode field, the normalized frequency V can be calculated by the following formula:

$$V = \frac{2\pi a}{\lambda} NA \tag{1}$$

, where a is the core radius, λ=846.5 nm is the wavelength of VCSEL, and NA is the numerical aperture of the fiber. The oxidation aperture of the VCSEL we use is about 7 μm, so a=3.5 μm. Then, by measuring the divergence angle of the beam, we can get NA of about 0.122. Therefore, we can get V=3.1694. The normalized transverse propagation constant U and W reflects the lateral distribution characteristics of the

light field in the core and cladding, respectively. They satisfy the following relationship.

$$V^2 = U^2 + W^2 \tag{2}$$

The characteristic equation of the *LP* mode is as follows

$$\frac{J_l(U)}{UJ_{l+1}(U)} = \frac{K_l(W)}{WK_{l+1}(W)} \tag{3}$$

, where J_l denotes the first-kind Bessel function of order l and K_l denotes second-kind modified Bessel function of order l. The relationship U, $W<V$ can be derived from Equation (2). By solving Equation (3), two sets of solutions (U=1.7992, W=2.6092 and U=2.7794, W=1.5233) corresponding to the LP_{01} and LP_{11} modes are obtained, proving that only these two modes can exist in the VCSELs used. This result is consistent with the experimental findings shown in Figure 2f in the main text. The LP_{01} mode corresponds to the fundamental mode in the VCSEL.

The normalized two-dimensional electric field distribution of the LP modes is as follows.

$$E(R_a) = \begin{cases} J_l(UR_a), & R_a \le 1 \\ J_l(U)\dfrac{K_l(WR_a)}{K_l(W)}, & R_a \ge 1 \end{cases} \tag{4}$$

, where $R_a = R/a$ is the normalized radius and R is the distance from the calculated point to the center of the core. By solving Equation (4), we can obtain the two-dimensional electric field distribution of the LP_{01} mode in the VCSELs.

Supplementary Note 2: Wave analysis of the individually coherent-mutually incoherent VCSEL array

In our experiment, VCSELs operate independently, emitting without mutual coherence. Consequently, even over relatively short averaging periods, the intensity distributions of individual emitters can be summed without the need to consider interference effects. For each VCSEL, due to its LP_{01} mode operation and narrow linewidth, it exhibits good coherence. Therefore, when calculating the far field of the VCSEL array, we first compute the far complex field of each VCSEL individually on the imaging plane based on angular spectrum diffraction, and then superimpose their amplitudes.

The VCSEL array consists of 64 units arranged in an 8×8 grid, with a pitch of 100 μm. Each VCSEL's light field is calculated independently and labelled from i=1 to 64. The light fields are set to a resolution of 1000×1000 pixels, with each pixel covering an area of 1 μm². By assigning the calculated LP_{01} mode field to the corresponding VCSEL positions, Supplementary Figure 4 is generated. It is assumed that the phase of each pixel in these light fields is uniform.

Based on angular spectrum diffraction[S2], the input complex-amplitude field from the n-th neuron located at (x_n, y_n) in the l-th layer to the n-th neuron in the (l+1)-th layer can be expressed as:

$$u_n^{l+1}(x,y) = F^{-1}\{F\{u_n^l(x_n,y_n)t_n^l(x_n,y_n)\}H(f_x,f_y)\} \quad (5)$$

where F and F^{-1} denote Fourier transform and inverse Fourier transform, respectively. The transmittance coefficient t represents the amplitude and/or phase modulation of a neuron and can be expressed as $t_n^l(x_n,y_n) = A_n^l(x_n,y_n)exp\,(j\varphi_n^l(x_n,y_n))$, where A and φ are amplitude and phase modulation of the neuron, respectively, and $j=\sqrt{-1}$. The diffractive neural networks we used are phase type, therefore, in our experiment, A was supposed to be a constant. $H(f_x,f_y)$ is the transfer function in spatial frequency domain and denotes the phase

delay after propagating a distance of *d*,

$$H(f_x, f_y) = exp\left((j2\pi d/\lambda)\sqrt{1-(\lambda f_x)^2-(\lambda f_y)^2}\right) \quad (6)$$

, where f_x and f_y are the spatial frequencies in the x and y direction, respectively and λ is the wavelength of incident light. Therefore, by superimposing the complex-amplitude field propagated from all the neurons in the *l*-th layer, the total complex-amplitude field on the (*l*+1)-th layer can be obtained.

Finally, by linear superimposing the output intensities of all the VCSELs, we can obtain the intensity distribution on the output layer, which can be represented by the following equation

$$\sum_{i=1}^{64} n_i * |u_i|^2 \quad (7)$$

, where n_i=0 or 1 represents whether the *i*-th VCSEL is on or off and $|u_i|$ is the amplitude distribution of the corresponding VCSEL on the output layer.

Supplementary Note 3: Training of the diffractive computing networks (MI-DNNs)

As shown in Supplementary Figure 6, we developed a forward propagation model for the MI-DNNs, using the light field calculation method outlined in Supplementary Note 2. For image recognition tasks, the loss function was defined by the cross-entropy error between the output and the target. For image denoising, precise control of the output light amplitude field was necessary, so we employed mean squared error (MSE) between the output and the target as the loss function. Then, we employed the stochastic gradient descent algorithm to back-propagate the errors and update the phase distribution in each layer to minimize the loss function. The training was performed using Google's TensorFlow 2.6.0 framework. The two diffractive layers were integrated on a commercial quartz plate. Therefore, the spacing between the two layers was set to 1 mm, corresponding to the thickness of the quartz plate. The MI-DNN chip was placed directly above the VCSEL lasers. Due to the presence of the driving circuit chip as a supporting layer, the distance from the emitting surface of VCSELs to the first diffractive layer is fixed to 0.963 mm, which is measured by a step meter. The distance from the MI-DNN to the output layer was set to 5 mm, which was determined to be the optimal value after testing various configurations (Supplementary Figure 9). Considering the fabrication tolerances of double-sided lithography and etching, the neuron size of the MI-DNN was set to 5 μm × 5 μm. The experiment demonstrated two tasks: image classification and image processing, with varying training sets and program parameters for each task. The obtained phase distributions of all the MI-DNNs are illustrated in Supplementary Figure 19.

a. Handwritten digital/letter recognition

For handwritten digit recognition, we used the Modified National Institute of Standards and Technology (MNIST) database. The original 28×28-pixel images were resized to binary 8×8-pixel images, matching the unit count of the VCSEL array. To recognize the digits 0 and 1, we trained the MI-DNN on 1000 images (500 per digit).

When recognizing all ten digits (0–9), we observed that some digital images lost key features upon resizing to 8×8 pixels, resulting in distorted images. Given the negative impact of these training images on the MI-DNN performance, we limited the training set to 1000 images (100 per digit). For 4-class handwritten letter recognition, we used the Extended MNIST Letters database. We selected four letters (“c,” “o,” “p,” and “s”), with 200 images per letter, as they usually retained distinguishing features after resized to 8×8 pixels. Here, according to the area of the VCSEL array, each layer of the MI-DNN was designed to have 200×200 neurons. In the output layer, multiple detection regions were defined, each corresponding to a recognized category, with light intensities indicating the inferred probabilities.

b. Image edge extraction

Unlike traditional convolutional kernels that use mathematical operations to extract edges, we trained the MI-DNNs to recognize specific edge patterns, effectively replacing the need for a 3×3 kernel. Therefore, we could use a kernel with 2×2 pixels, corresponding to 2×2 VCSELs. As with digit recognition, the output layer included two detector regions, corresponding to TRUE and FALSE. The MI-DNNs identified the target image as TRUE only when the edge pattern was present. When all 2×2 VCSELs are off, no light is emitted, so this scenario is excluded from the training set, and the output is directly defined as FALSE during testing. Each layer of the MI-DNNs for this task contained 100×100 neurons.

c. Image denoising

Taking the advantage of inherent mutual incoherence of VCSELs, we can independently control the output of each VCSEL channel, enabling the multiplication function of convolution. The outputs from all VCSELs are defined at the same region in the output layer to achieve the light intensity superposition, thereby enabling the addition function. Here, each layer of the MI-DNN contained 120×120 neurons. During training, we employed a multi-objective optimization approach, constructing 9

MSE losses corresponding to the VCSELs respectively, which were summed to form the total loss defined as follows,

$$L_{total} = \frac{1}{n}\sum_{i=1}^{9} (a_i A - A_i)^2 \tag{8}$$

where A denotes the target amplitude in the detection area, A_i denotes the output amplitudes of each VCSEL, and a_i denotes the corresponding filter coefficient. According to the linear relationship between light intensity and square of amplitude, the values of a_i is set as follows.

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.707 & 0.5 \\ 0.707 & 1 & 0.707 \\ 0.5 & 0.707 & 0.5 \end{bmatrix}$$

Supplementary Note 4: Light energy per frame and diffraction efficiency and signal-to-noise ratio (*SNR*) of the OGPU under 25 MHz high-speed testing

Under direct current (DC) operation, the optical intensity of a single VCSEL at 1.7 V with an injection current of approximately 2 mA is measured to be 1.27 mW (Fig. 2b). During high-speed operation at a frequency of 1 MHz, the VCSEL current remains stable for a certain duration during lasing. The average signal detected by the high-speed photodetector during this period is 15.36 mV (Supplementary Figure 14).

At an operating frequency of 25 MHz, the VCSEL has a pulse width of approximately 2.05 ns, with a corresponding average photodetector signal of 1.85 mV (Supplementary Figure 14). Assuming that the optical intensity at 1 MHz remains consistent with that observed under DC operation, the light energy generated by a single VCSEL per frame at 25 MHz can be calculated as:

$$1.85\ \text{mV}/15.36\ \text{mV}*1.27\ \text{mW}*2.05\ \text{ns} = 3.136*10^2\ \text{fJ}.$$

Then, when encoding the handwritten digit “0”, an average number of 23.65 VCSELs are lit per frame, whereas for the handwritten digit “1”, an average number of 11.22 VCSELs are lit per frame. Therefore, the average light energy per frame can be calculated as:

$$3.136*10^2\ \text{fJ}*23.65 = 7.42\ \text{pJ (For digit "0")},$$

$$3.136*10^2\ \text{fJ}*11.22 = 3.52\ \text{pJ (For digit "1")}.$$

Diffraction efficiency is determined by dividing the light intensity in the target detection region by the total input light intensity. For the 2-class MNIST classification task, when the input image is labelled “1”, the high-speed photodetector records an average peak signal of 10.8 mV in the corresponding detection area for correctly recognized instances (Fig. 3d). For images labelled “0”, the detected average peak signal is 9.4 mV (Fig. 3d). Since a single VCSEL illuminating the detector produces a peak signal of 3.7 mV (Supplementary Figure 14), the average diffraction efficiency can be calculated as:

$$9.4\ \text{mV}/(3.7\ \text{mV}*23.65) = 10.74\%\ \text{(For digit "0")},$$

10.8 mV/(3.7 mV*11.22) = 26.02% (For digit "1").

The *SNR* of the MI-DNN is calculated by the following equation

$$SNR = 10log_{10}(P_s/P_n) \quad (9),$$

where *P*s and *P*n are the signal strength and noise signal strength of the high-speed detector in the target detection area, respectively. During the test, the average noise of the detector is about *P*n=0.428 mV. Therefore, the *SNR* of the OGPU is

$10log_{10}(9.4\ \text{mV}/0.428\ \text{mV})$ = 13.42 dB (For digit "0"),

$10log_{10}(10.8\ \text{mV}/0.428\ \text{mV})$ = 14.02 dB (For digit "1").

Supplementary Note 5: Computing performance of the OGPU

Due to the mutual incoherence properties of VCSELs, the computational power of the OGPU can be determined by calculating the computing capacity of each VCSEL channel individually and then multiplying it by the total number of VCSEL units. The computational power of the diffractive networks is primarily dependent on the number of nodes between layers. Following pioneering research[3], and given the absence of nonlinear activation between the two diffractive layers in the MI-DNN, we treat the two layers as a single effective diffractive layer with an equivalent propagation distance (about 1473 μm) equal to their average. Fig. s1 below shows the diffraction pattern of a VCSEL on the on the effective diffractive layer. The size of the diffractive neurons in the MI-DNN chip is 5×5 μm^2. Thus, the corresponding number of nodes in this layer is calculated as $k_1 = \pi \times (374\ \mu m/2)^2 / (5\ \mu m \times 5\ \mu m) \approx 4{,}392$. Thus, the total number of operations performed by a VCSEL during the diffraction process is $k_1 \times n$, where n is the number of detection regions on the output layer, equivalent to the number of classifications in the MI-DNNs. Additionally, as the light field passes through the diffractive layers, each layer modulates the light field, contributing further operations, quantified as k_1.

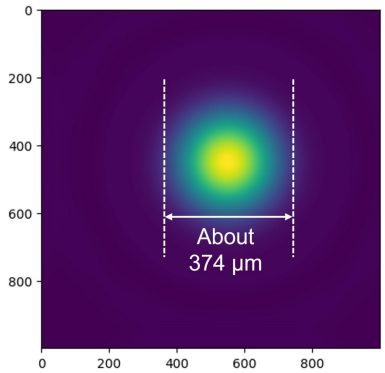


Fig. s1. Light field pattern of a single VCSEL unit on the effective diffractive layer.

Therefore, the total number of operations for a VCSEL channel in a single computation is given by

$$n_{ops}(\text{a VCSEL}) = k_1 \times n + k_1 \tag{10}$$

Thus, with 64 VCSELs operating at 25 MHz under mutual incoherence (i.e.,

independent channels), the overall computational power of the OGPU with 64 VCSELs is expressed as:

$$n_{ops}(\mathrm{OGPU}) = 64 \times (k_1 \times n + k_1) \times f \tag{11}$$

where f is the calculation frequence. Therefore, at the calculation frequency of 25 MHz, we can obtain the computing power under different tasks.

(1) 10-class MNIST dataset: $64\times(k_1\times10+k_1)\times25\times10^6=7.73\times10^{13}$ OPs=77.3 TOPS

(2) 4-class EMNIST dataset: $64\times(k_1\times4+k_1)\times25\times10^6=3.51\times10^{13}$ OPs=35.1 TOPS

(3) 2-class MNIST dataset: $64\times(k_1\times2+k_1)\times25\times10^6=2.11\times10^{13}$ OPs=21.1 TOPS.

The energy consumption $E(\mathrm{OGPU})$ of the OGPU is primarily attributed to three components: the VCSELs, the FPGA chip (Xilinx Artix7 XC7A35T), the driver board (equipped with 4 TI LVC16244 chips) and the high-speed detector (Thorlabs, DET025AL, 2GHz). We assume that the VCSEL operating at 25 MHz maintains the same power consumption as in DC operation (though the actual value is expected to be lower). At this frequency, each switching cycle lasts 40 ns, with a pulse width of approximately 2 ns. Therefore, $E(\mathrm{VCSELs}) = 1.7\ V \times 2\ mA \times 64\ units \times 2\ ns/40\ ns = 10.88$ mW, $E(\mathrm{FPGA}) \approx 69$ mW, $E(\mathrm{driver\ board}) = 72\ \mu\mathrm{W} \times 4 = 0.288$ mW and $E(\mathrm{photodetector}) = 1.2$ mW.

Thus, the energy efficiency η of OGPU can be calculated as

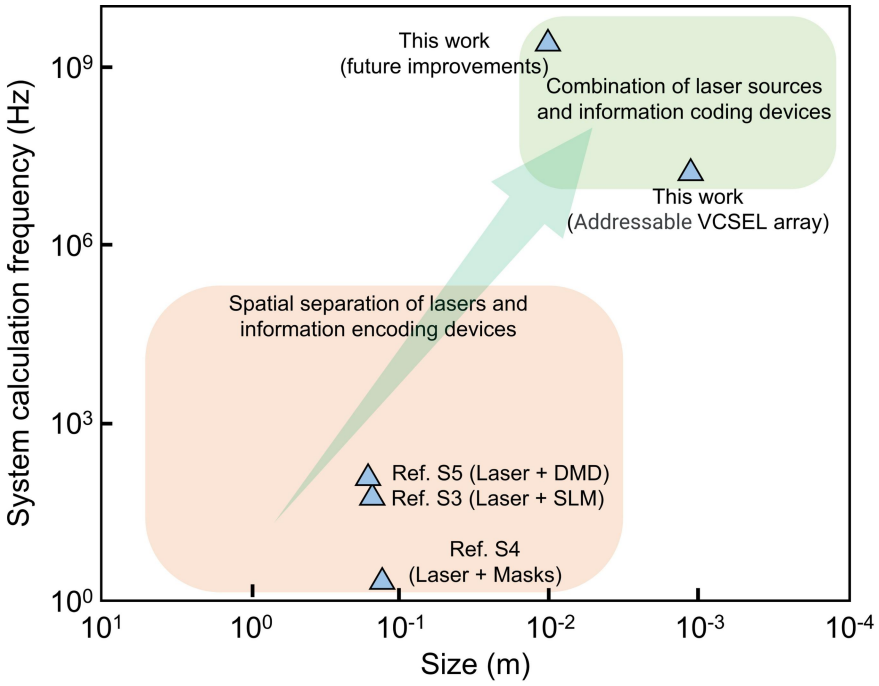
$$\eta = \frac{n_{ops}(\mathrm{OGPU})}{E(\mathrm{OGPU})} \tag{12}$$

We can obtain the energy efficiency under different tasks.

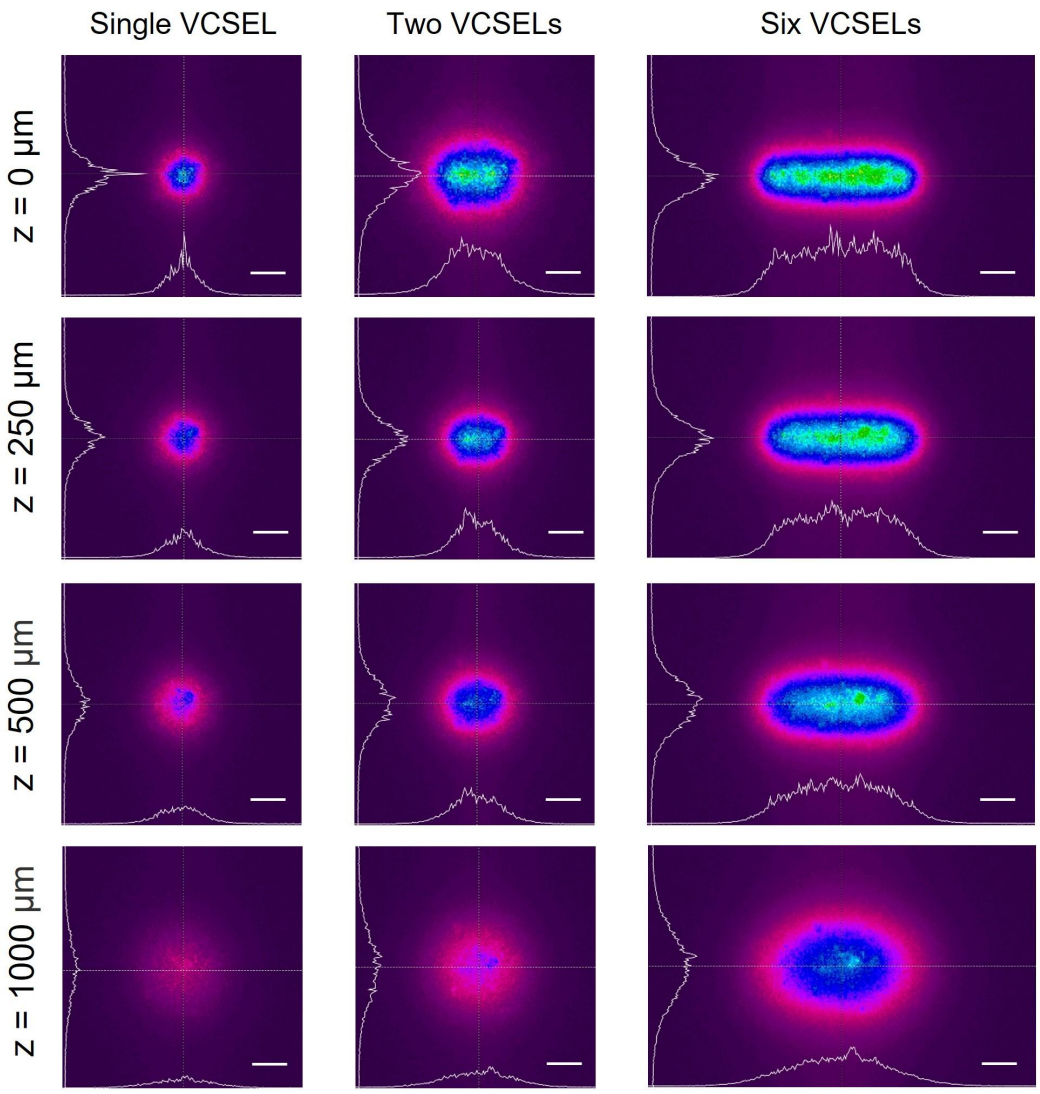
(4) 10-class MNIST dataset: 77.3 TOPS / 81.37 mW=9.50×10^2 TOPS/W

(5) 4-class EMNIST dataset: 35.1 TOPS / 81.37 mW=4.31×10^2 TOPS/W

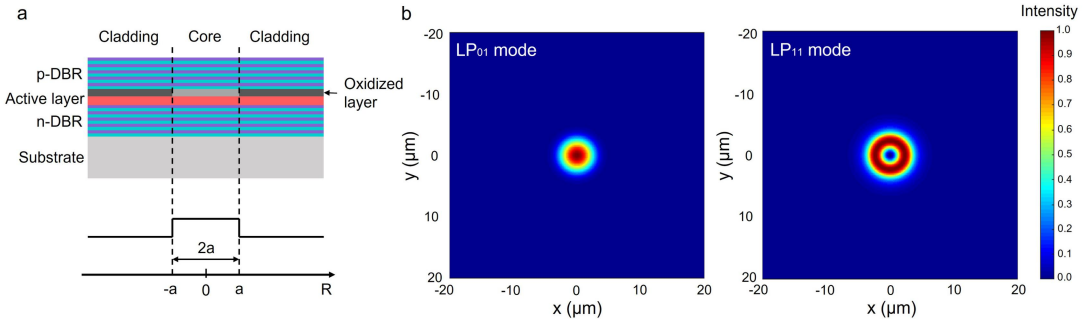
(6) class MNIST dataset: 21.1 TOPS /81.37 mW=2.59×10^2 TOPS/W



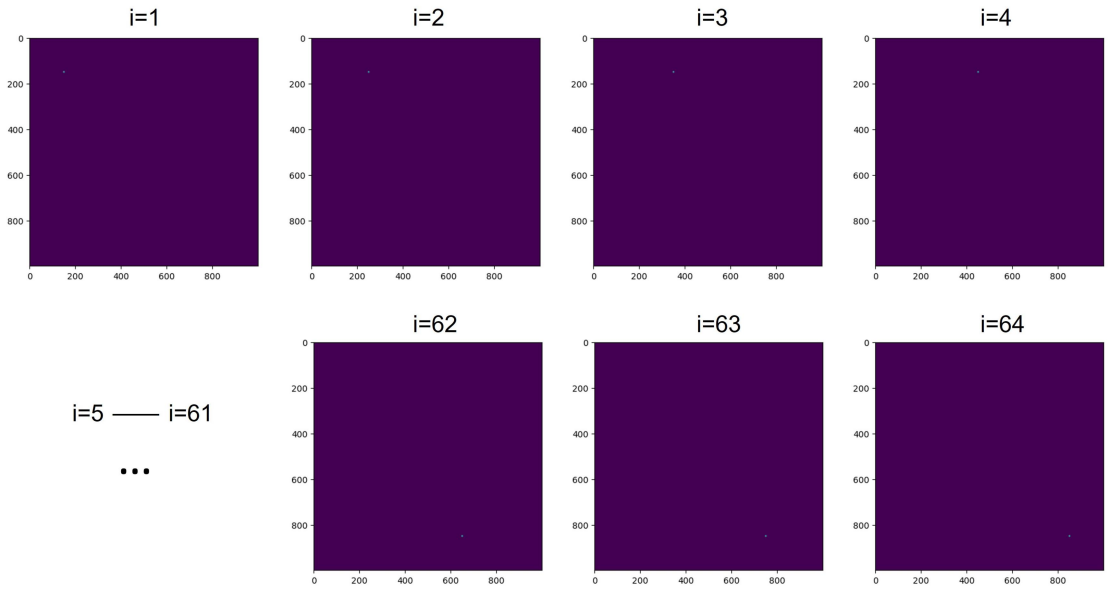
Supplementary Figure 1. Comparison of this work with other diffractive computing approaches, focusing on the size of input devices (laser sources and information encoding devices) and computational frequency. The addressable VCSEL array, as an on-chip active device, combines the functions of traditional laser sources and information encoding devices, such as spatial light modulators (SLMs) and digital micromirror devices (DMDs). This integration significantly enhances both the computational frequency and the system's overall compactness.



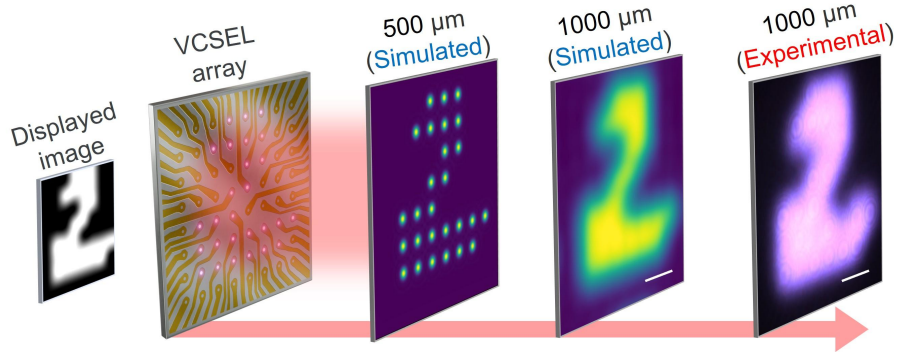
Supplementary Figure 2. Light field patterns at varying propagation distances with different numbers of VCSELs activated. No interference is observed between the VCSELs, confirming the mutual incoherence between the units. Scale bar: 100 µm.



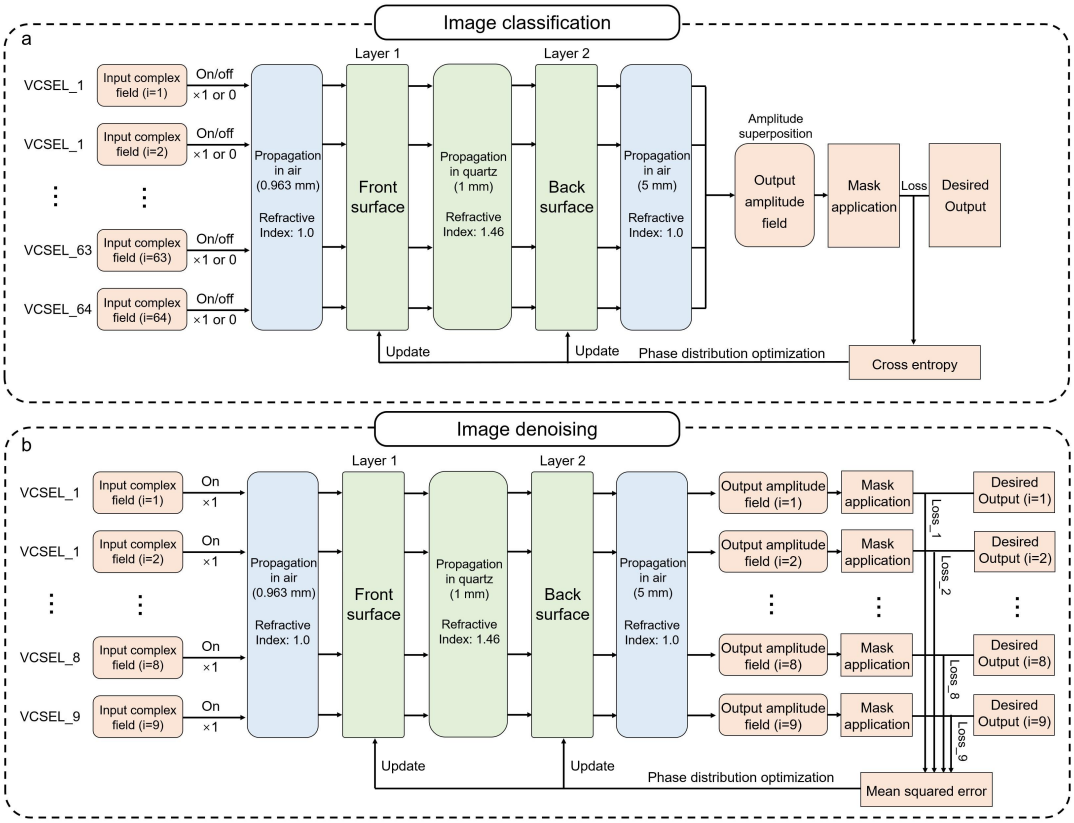
Supplementary Figure 3. Calculation of guided modes in the VCSEL. a. Diagram of the epitaxial structure of the VCSEL, showing the refractive indices of the core and cladding regions. **b.** Calculated amplitude fields of the guided LP_{01} and LP_{11} modes in the VCSEL.



Supplementary Figure 4. Input amplitude field of each VCSEL in the 8×8 array at the input layer. The input layer consists of 1000×1000 pixels, with each pixel measuring 1 μm×1 μm.

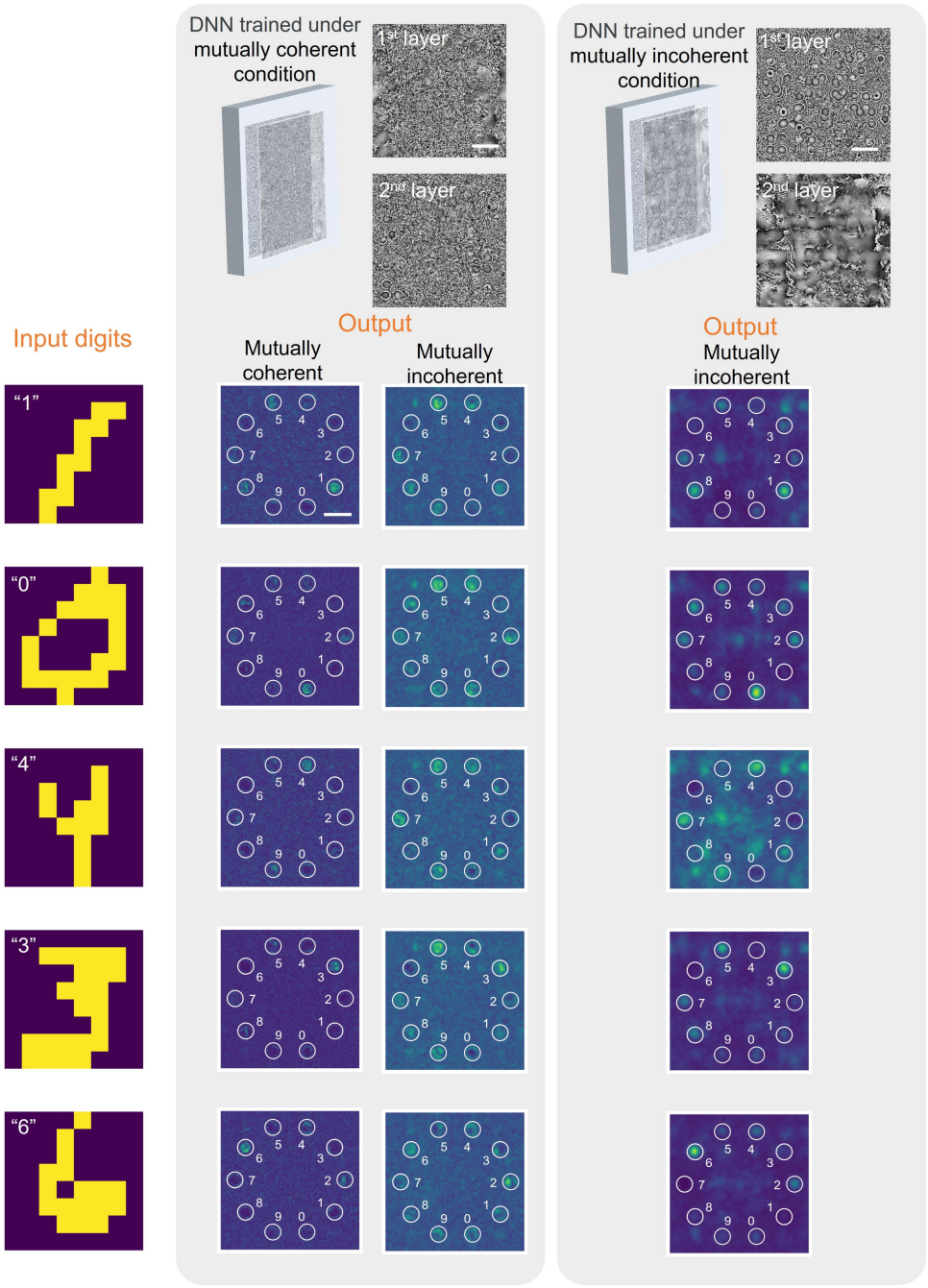


Supplementary Figure 5. Calculated and experimentally measured far-field patterns at varying propagation distances of the VCSEL array displaying the image of "2". Scale bar: 200 μm.

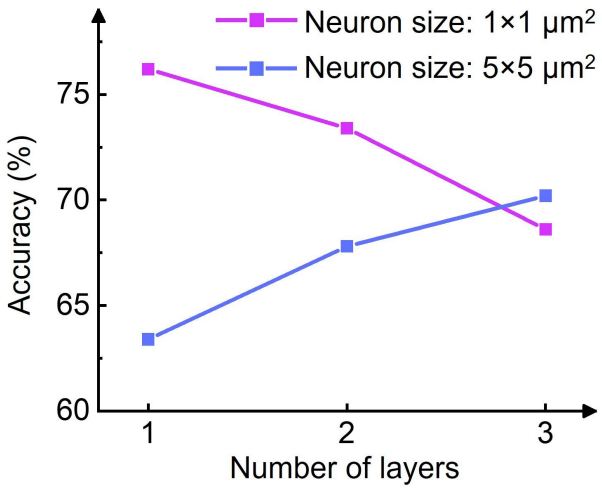


Supplementary Figure 6. Training flowcharts of the MI-DNNs for different tasks.

a. Image classification. **b.** Image denoising.

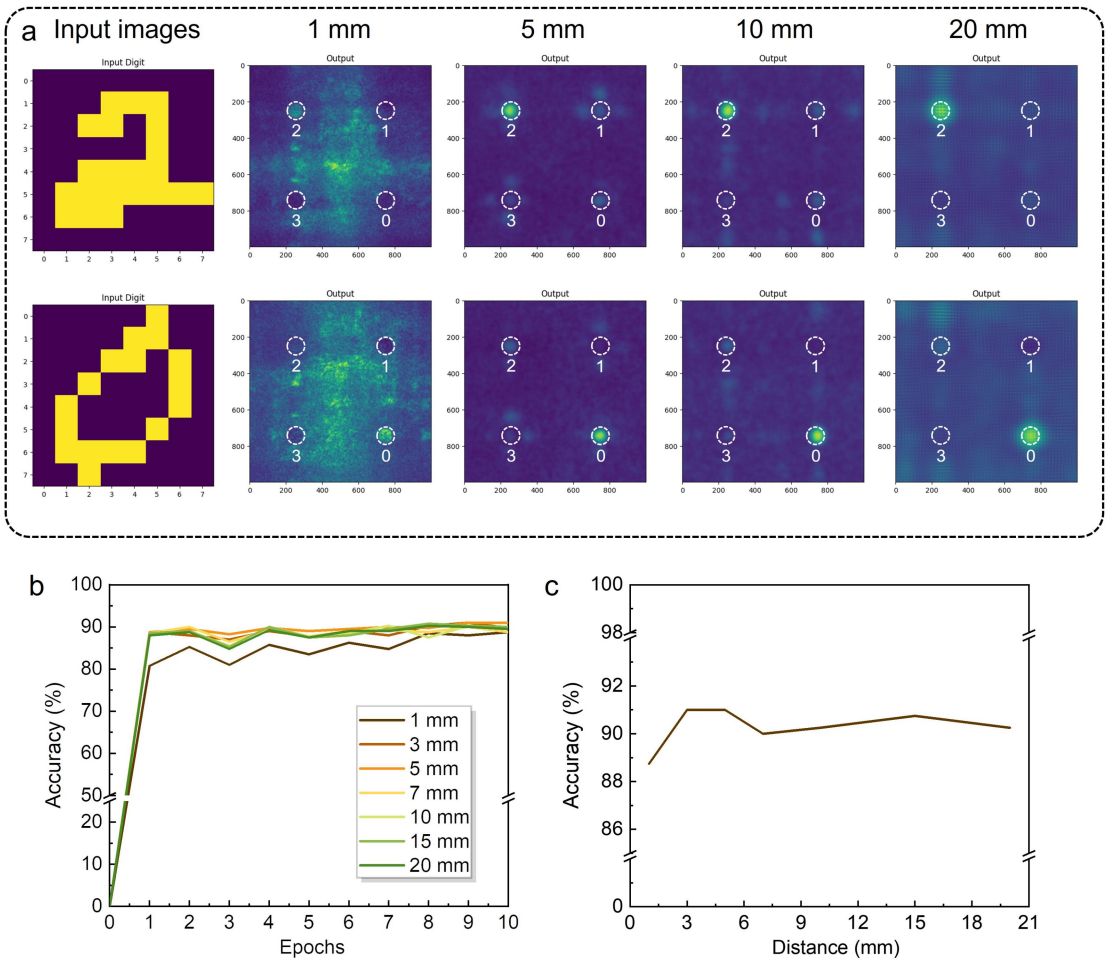


Supplementary Figure 7. Examples of the outputs of the MI-DNNs trained under coherent and mutually incoherent conditions. When the VCSEL units are assumed to be coherent, the trained MI-DNN correctly recognizes the input digital images. However, when the light source is switched to mutually incoherent VCSELs, as in the actual scenario, the MI-DNN misrecognizes the same input images. In contrast, the MI-DNN trained with mutual incoherence between VCSEL units correctly identifies the images. Scale bar: 200 μm.

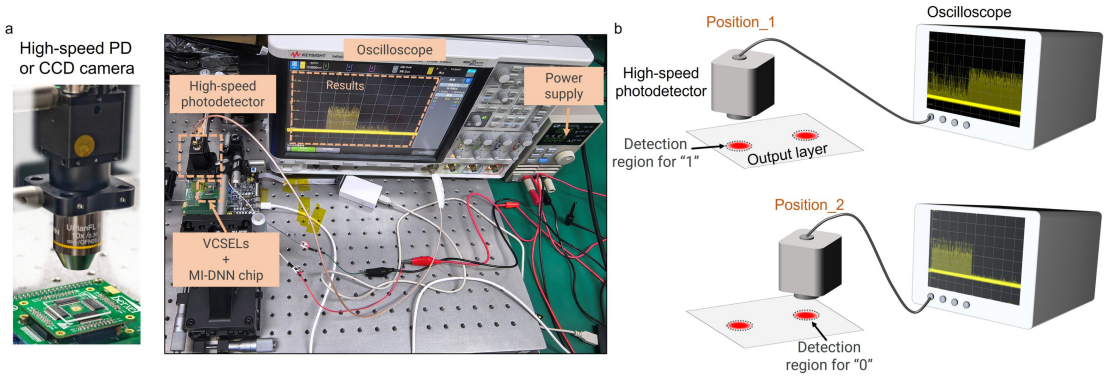


Supplementary Figure 8. Accuracy of the MI-DNN versus the number of diffractive layers for different neuron sizes. Here, the MI-DNNs are trained using the following parameter: the size of the diffractive layer is fixed at 1000×1000 μm^2. The distance from the input layer to the DNN is 500 μm, the distance from the output layer to the DNN is 1000 μm, and the DNN size is 1000×1000 μm^2 with a neuron size of 1 μm^2. The layer spacing of multi-layer DNN is 200 μm.

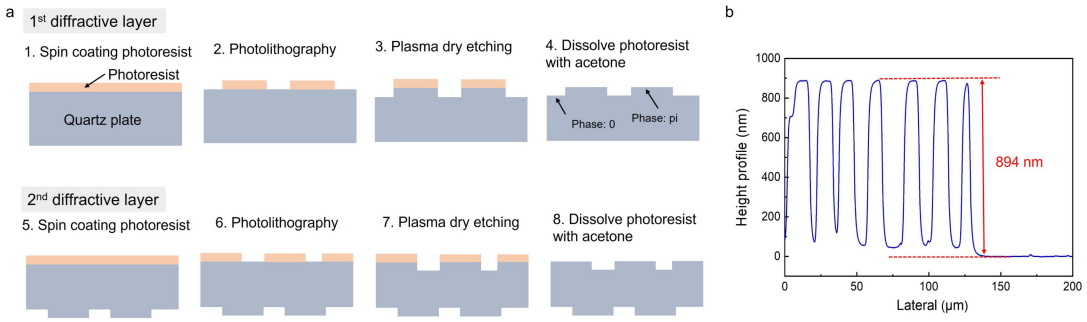
From the results, it can be seen that with sufficient spatial bandwidth product (1 μm^2 neuron size), and a single-layer MI-DNN achieves optimal training results. When the neuron size is increased to 25 μm^2, the spatial bandwidth product of a single layer decreases by 25 times; in this case, adding layers can effectively increase the spatial bandwidth product in space and improve training accuracy. In the experiment, considering the limitations of photolithography precision, we selected a neuron size of 25 μm^2 and enhanced the space-bandwidth product by implementing a double-layer MI-DNN.



Supplementary Figure 9. Distance optimization from the MI-DNN chip to the output layer. a. Outputs of the MI-DNNs trained at various output layer distances for 4-class MNIST classification. When the output layer distance is either too small or too large, the MI-DNN fails to efficiently modulate the output light field, leading to reduced diffraction efficiency. **b.** Training accuracy vs. training epoch curve for MI-DNNs at different output layer distances. **c.** Maximum accuracy achieved by the MI-DNNs within 10 training epochs as a function of output layer distance.



Supplementary Figure 10. Separated optical setup for high-speed test. a. Experimental setup of the test system showing the integrated architecture of VCSELs and MI-DNNs within the OGPU. PD: photodetector; CCD: charge-coupled device. **b.** Schematic of using a high-speed detector to collect light intensity signals from two detection regions. Detection is performed by moving the spatial position of the high-speed detector to each region.

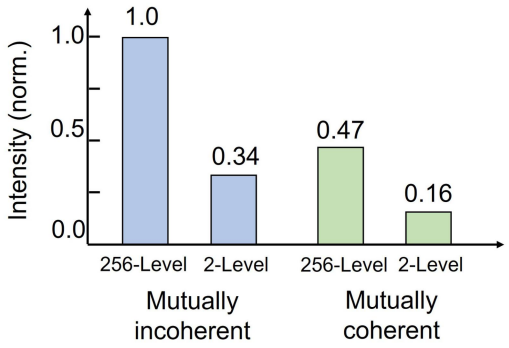


Supplementary Figure 11. Fabrication of the MI-DNN chip with monolithic integration of the two diffractive layers. a. Schematic of the fabrication process. **b.** Height profile of the diffractive layers within a 200 μm range. The average etching depth achieved by dry etching is approximately 894 nm.

a

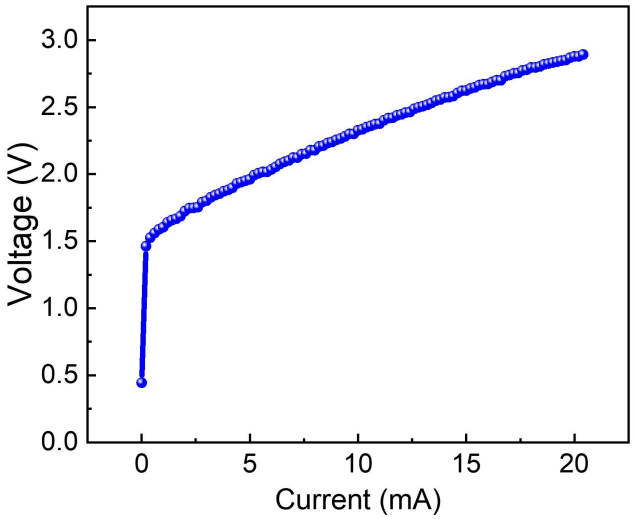
	Accuracy	
	256-Level	2-Level
4-class EMNIST	87.75%	87.75%
10-class MNIST	78.2%	76.8%

b

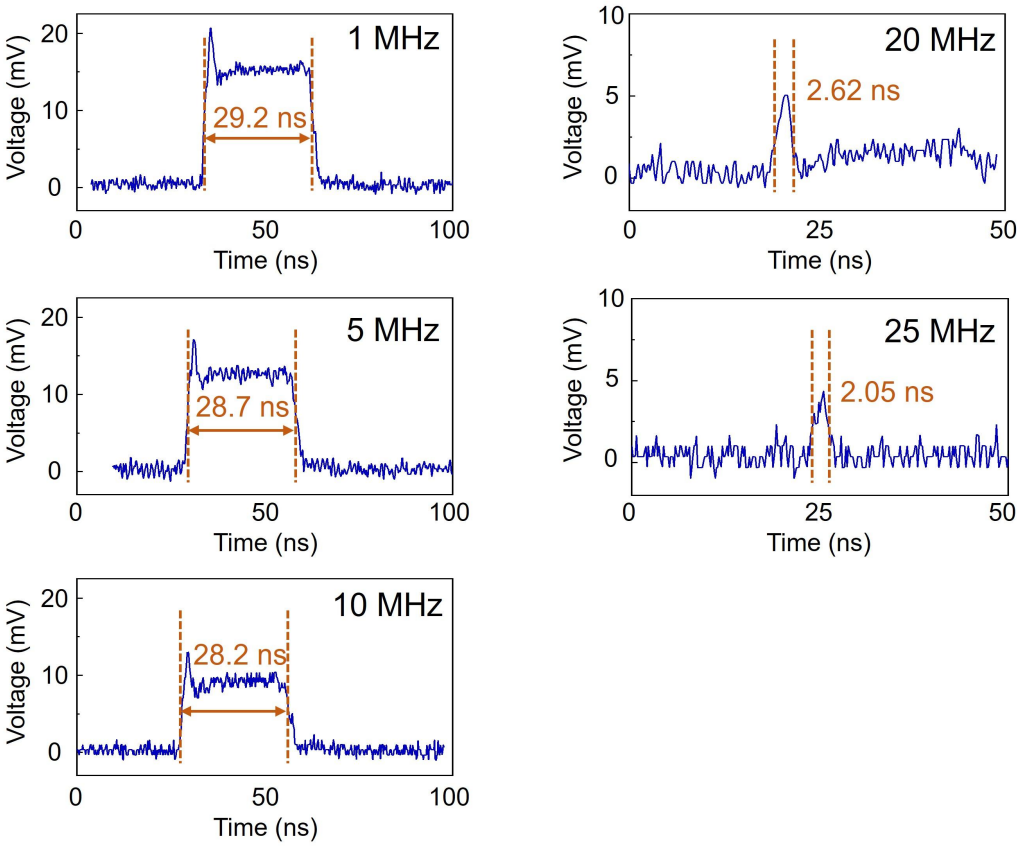


Supplementary Figure 12. Effect of phase binarization on the MI-DNNs' performance. a. Accuracies of the MI-DNNs before and after phase binarization. The results indicate that phase binarization has little impact on classification performance.

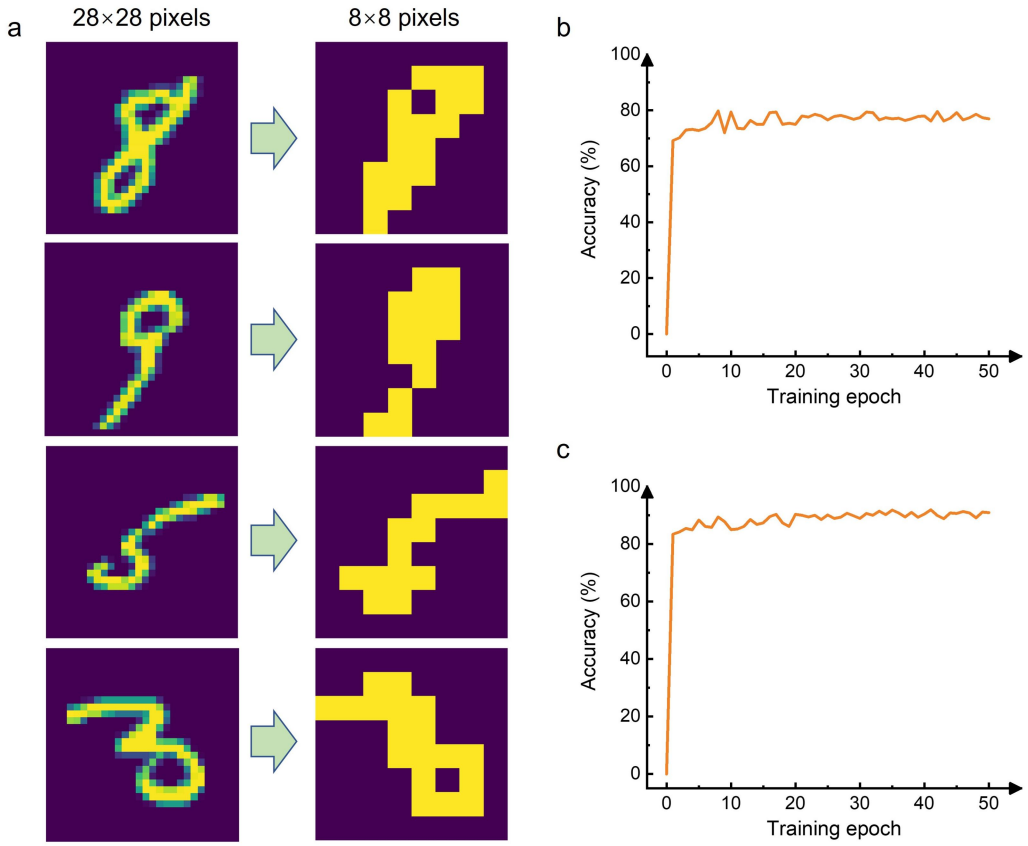
b. Average intensity in the target detection area of the output layer for correctly recognized training set images before and after phase binarization. The diffraction efficiency of DNNs decreases after phase binarization, leading to a reduction in diffraction efficiency. However, MI-DNNs still maintain a higher diffraction efficiency compared to DNNs trained with coherent light.



Supplementary Figure 13. Voltage-current characteristic curve of the VCSELs. At a working voltage of 1.7 V, the VCSELs operate at an injection current of approximately 2 mA.

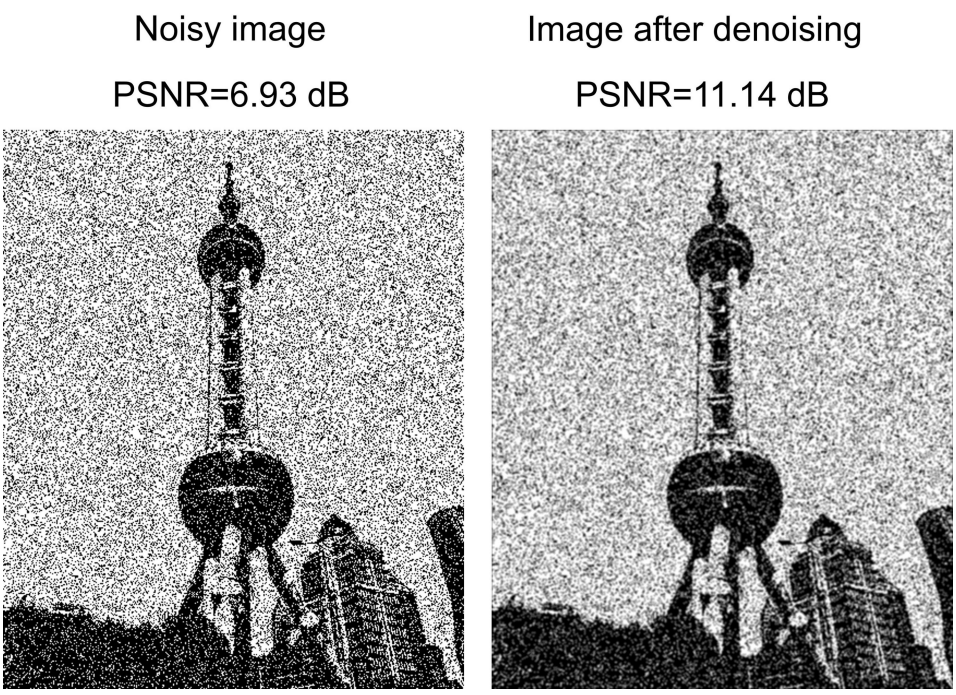


Supplementary Figure 14. Pulse widths of the a VCSEL under different modulation bandwidths.

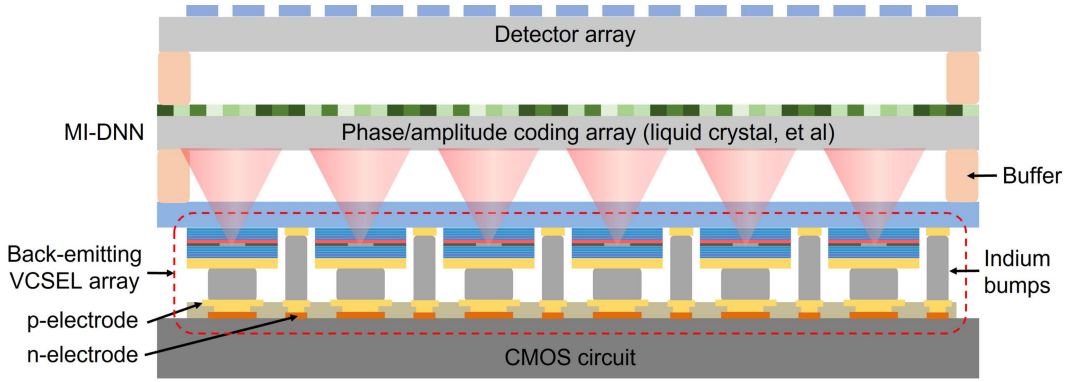


Supplementary Figure 15. Impact of VCSEL array resolution on MI-DNN classification accuracy. a. Representative examples illustrating the effect of downscaling 28×28-pixel handwritten digit images to 8×8 pixels, where resolution loss leads to the weakening or loss of critical label features. **b.** Training accuracy versus epoch curve for a conventional coherent-light-based diffractive neural network (DNN) using 8×8-pixel images for 10-class classification. **c.** Training accuracy versus epoch curve for a mutually incoherent DNN (MI-DNN) using full-resolution 28×28-pixel images for the same classification task.

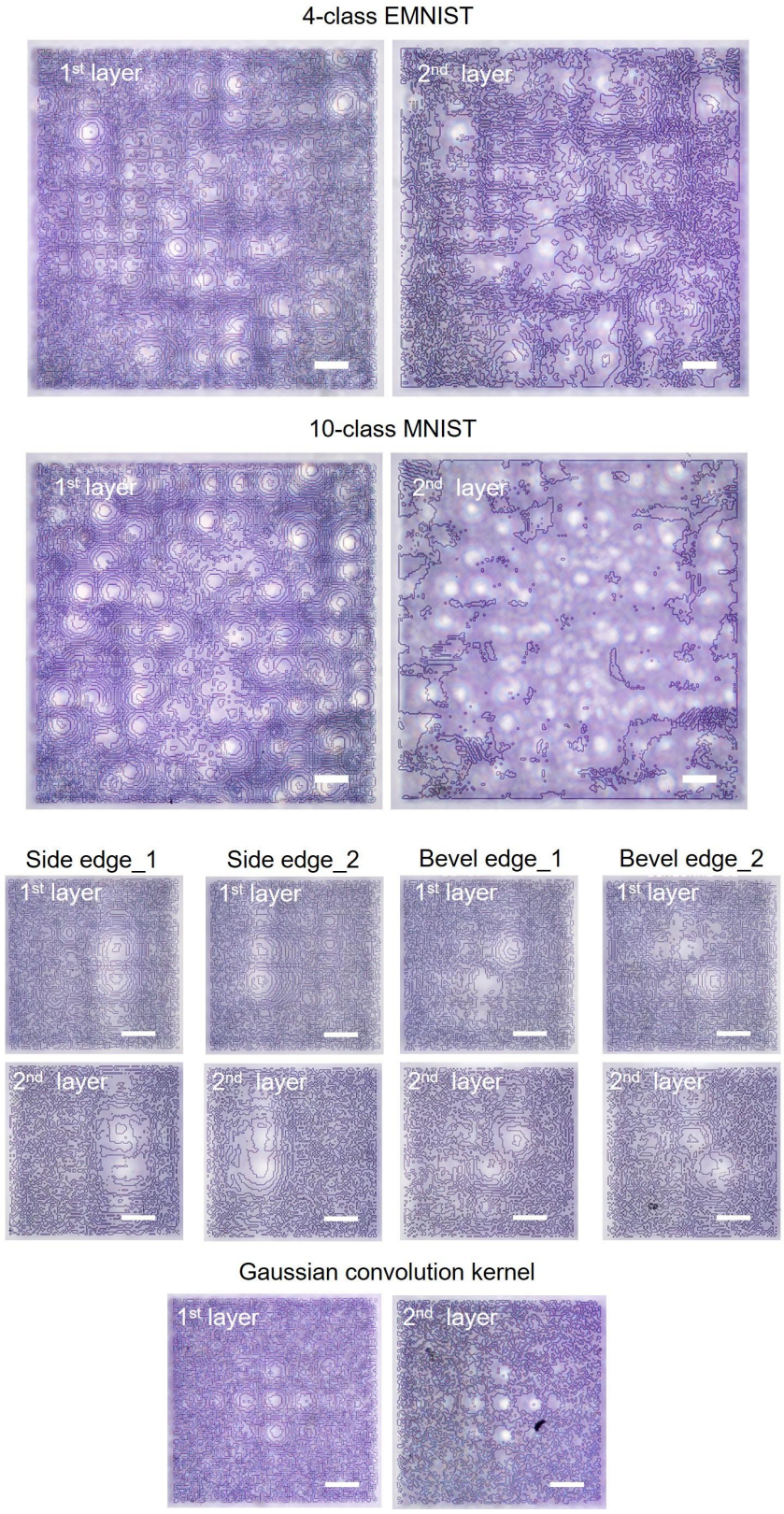
Supplementary Figure 16. Edge extraction results from various images processed by the four MI-DNNs, each with different target edge patterns. The grayscale of the images has been inverted for clearer observation.



Supplementary Figure 17. Images before and after denoising using the MI-DNN-based Gaussian filtering. The noisy image has a PSNR of 6.93 dB, which increases to 11.14 dB after denoising.



Supplementary Figure 18. Vision for the future OGPU architecture based on a large-scale VCSEL array. a. The OGPU utilizes substrate-emitting VCSELs as light sources and achieves large-scale, high-speed addressing through a specialized CMOS chip. The diffractive layers incorporate reconfigurable materials, such as liquid crystals, to enable programmable functions. All components of the OGPU are vertically on-chip integrated.



Supplementary Figure 19. Optical images of the two diffractive layers of the obtained MI-DNNs. Scale bar: 100 μm.

References

S1. Okamoto, K. *Fundamentals of Optical Waveguides* (Second Edition). Academic Press, 57-158 (2006).

S2. Goodman JW. *Introduction to Fourier optics*. Roberts and Company publishers (2005).

S3. Chen Y.*, et al.* All-analog photoelectronic chip for high-speed vision tasks. *Nature*, (2023).

S4. Luo X.*, et al.* Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible. *Light: Science & Applications* **11**, 158 (2022).

S5. Zhou T.*, et al.* Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics* **15**, 367–373 (2021).