Supplementary Materials for

Highly Scalable Machine Vision Enabled by Meta-Optics-based Ultra-Wide Neural Network

Mingcheng Luo[1], Meirui Jiang[2], Bhavin J. Shastri[3], Nansen Zhou[4], Wenfei Guo[1], Jianmin Xiong[1], Dongliang Wang[1], Renjie Zhou[4], Chester Shu[1], Qi Dou[2], Chaoran Huang[1*]

[1*]Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.
[3]Centre for Nanophotonics, Department of Physics, Engineering Physics & Astronomy, Queen's University, 64 Bader Lane, Kingston K7L3N6 ON, Canada.
[4]Department of Biomedical Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

*Corresponding author. Email: crhuang@ee.cuhk.edu.hk

This PDF file includes:

Figs. S1 to S22
Tables S1 to S8
References (1 to 53)

Contents

Note 17. COVID-19 detection based on Chest X-ray images

Note 18. Intracranial hemorrhage detection based on Brain Computed Tomography (CT) images

Note 19. Recognition of lung diseases based on Chest X-ray images

Note 20. Breast cancer detection based on whole slide images

Note 21. Human action recognition based on video streams

Note 22. Digital neural networks setup and training for SAM, ViT, ResNet-50 model

Note 23. Time consumption of the MOLM during training and inference phases

Note 24. Energy consumption of the MOLM during training and inference phases

Note 25. TSNE analysis for experimentally captured images

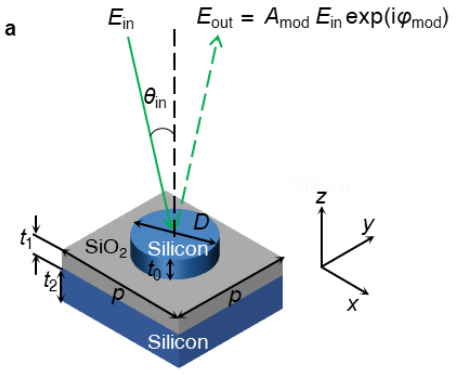
Note 26. Explanation of the enhancement of information entropy and high-frequency components

Note 27. Performance comparison tables

Note 1. FDTD simulations of the optical responses of the metasurface pillar

The proposed metasurface is created based on a silicon-on-insulator (SOI) substrate whose top silicon layer is the cylindric nanodisk, as shown in Figure S1a. The geometric parameters of the metasurface pillar in the simulations are listed in the table of Figure S1b. The thickness of the bottom silicon layer t_2 = 725 μm is in the fabricated metasurface device. We choose t_2 = 1 μm is in the simulations to save the computation costs.

A Finite-difference time-domain (FDTD) method is used to simulate the optical responses of the nanodisk. The light source is a plane wave source with a linear polarization along the *x*-axis. The boundaries at the *x*-*y* plane are periodic to emulate the periodic nanodisks of the metasurface, and the boundaries at the *z* plane are perfect match layers. The size of the mesh grids ranges from 0.25 nm to 8 nm. Figure S1c shows the phase and amplitude coefficients versus the diameters of the nanodisk under normal incidence. It can be found that both phase and amplitude can be modulated with a value determined by the nanodisk diameter.



b

Parameters	Values
p	500 nm
t_0	220 nm
t_1	2 μm
t_2	1 μm

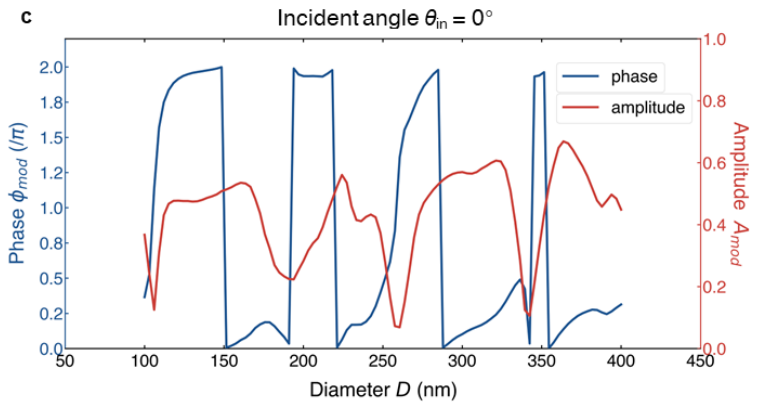


Figure S1. (a) The structural illustration of the metasurface pillar. (b) The summary table of the geometric parameters used in the simulations. (c) The simulated phase and amplitude modulation coefficients versus the radii of the pillar under normal incidence.

Note 2. Justification of the choice of the standard deviation (SD) of the phase modulation pattern of the optical metasurface

In this section, we will explain why choose a SD of 0.4 π for the phase distribution and how the SD of the amplitude distribution influences the system performance respectively.

(1) The choice of Gaussian random phase distribution with an SD of 0.4 π: We would like to clarify that this value was selected in consideration of fabrication limitations, rather than to suggest that 0.4π represents a globally optimal choice. FDTD simulations indicate that the achievable phase modulation spans by the meta-atom is the full range from 0 to 2π. However, near 2π, the relationship between pillar diameter and phase exhibits a plateau, as shown in Figure S2b. To ensure robustness against fabrication tolerances, we intentionally avoid this regime, since small fabrication deviations can cause the phase to collapse toward values near 2π. Consequently, when selecting the SD, we deliberately restrict it to 0.4π in order to minimize the number of diameter values that fall within the plateau region.

Furthermore, we simulated the system performance for standard deviations of 0.4 π and 0.2 π, respectively. The results indicate that the configuration with 0.4 π achieves superior performance. In this case, the classification accuracy reaches 99 %, which is already close to the upper limit. Therefore, we chose SD = 0.4π in our chip design. Our selection is based on empirical evaluation and may not be globally optimal.

(2) The influence of the SD of the amplitude distribution on the metasurface design: Since the phase and amplitude modulation coefficients are interdependent, only one can be independently designed. Once the diameter is set to achieve the desired phase modulation, the amplitude coefficient is correspondingly fixed and

is found to approximately follow a Gaussian distribution, as shown in Figure S2a. As a result, the complex-valued modulation implemented by the meta-atoms also follows a Gaussian distribution.

We have also conducted the simulations based on the MNIST dataset to investigate how the SD of phase distribution influences the system performance at different numbers of meta-atoms. Figure S2c shows that the MNIST accuracy is significantly promoted by increasing the number of meta-atoms for various phase distributions with an SD ranging from 0.2 π to 0.6 π. A larger SD contributes to the higher accuracy under various numbers of meta-atoms. However, when the number of meta-atoms increases to 4 million, the task performance at an SD of 0.4 π is very close to that at an SD of 0.6 π, with an accuracy of 96.5 % at an SD of 0.4 π and 96.8 % at an SD of 0.6 π. As the number of meta-atoms further increases to 41 million, the performance gap between an SD of 0.4 π and a higher SD could be further narrowed. The above results imply that an SD of 0.4 π in our design can achieve not only good fabrication tolerance but also superior performance.

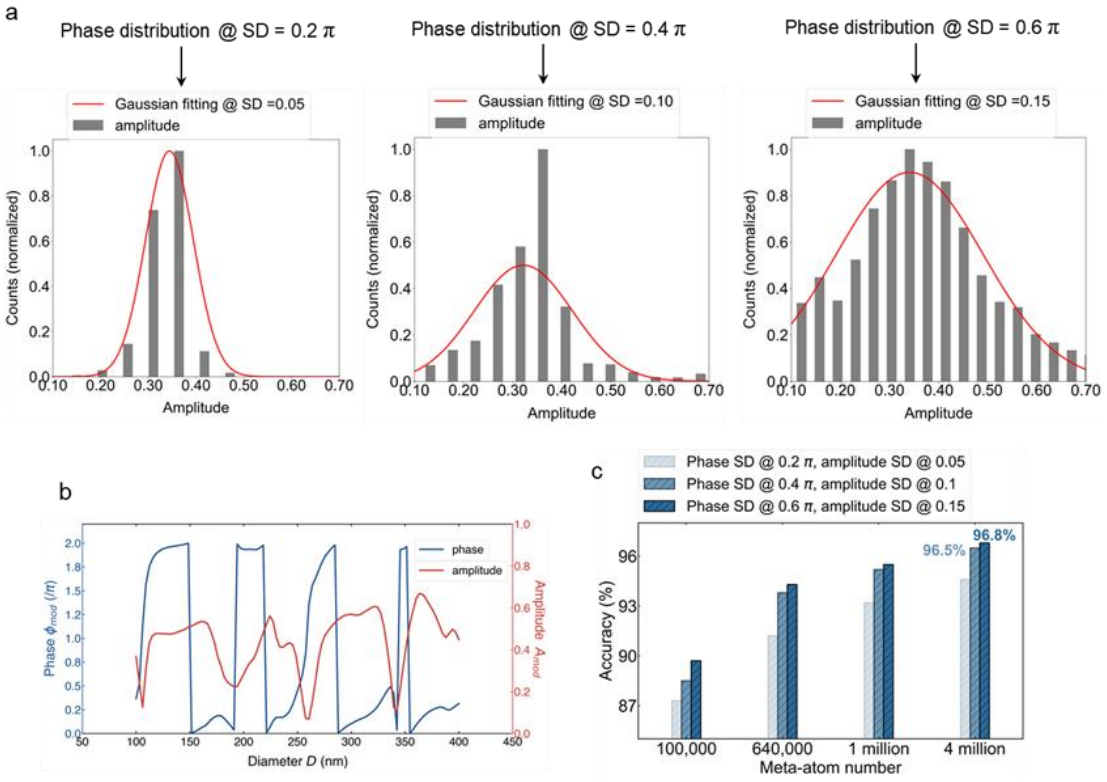


Figure S2. (a) Illustration of the statistical distribution of the amplitude pattern determined by the phase distribution with various SD from 0.2 π to 0.6 π. (b) The mapping relationship between the phase and amplitude

modulation coefficients and the meta-atom diameter. (c) The simulated MNIST accuracy versus the number of meta-atoms for different SD of the phase pattern.

Note 3. Fabrication process of the metasurface chip

The metasurface chip is fabricated based on a multiple-wafer-project (MPW) process offered by a commercial electron-beam-lithography-based silicon foundry. The fabrication process starts with a silicon-on-insulator (SOI) wafer. The thickness of the device layer is 220 nm, the thickness of the buffer oxide layer is 2 μm, and the thickness of the silicon substrate is 725 μm. A layer of electron-sensitive photoresist is coated on the device layer of the wafer, followed by heating to strengthen the photoresist. After that, a 100 keV electron gun is used to define the nanodisk patterns of the metasurface. The wafer is then chemically developed, and the patterned photoresist remains on the substrate. Finally, an anisotropic RIE etching process is performed to transfer the pattern of the photoresist into the silicon device layer. The etching process will continue until the surface of the buffer oxide layer is exposed.

Note 4. Mathematical model and experimental setup of the metasurface-based optical learning machine (MOLM)

1. Mathematical model

Firstly, we use a combination of the phase-only spatial light modulator (SLM) and two linear polarizers to generate the input optical image $\mathbf{U}_0$, as introduced in the reference [1]. The relationship between the SLM phase pattern $\boldsymbol{\varphi}_{\mathrm{SLM}}$ and the generated optical image $\mathbf{U}_0$ can be expressed by the following equation:

$$\mathbf{U}_0 = \frac{1}{2} A_0(\exp(i\boldsymbol{\varphi}_{\mathrm{SLM}}) + 1) = A_0 cos(\frac{\boldsymbol{\varphi}_{\mathrm{SLM}}}{2}) \exp\left(\frac{i\boldsymbol{\varphi}_{\mathrm{SLM}}}{2}\right)$$

where A_0 is the amplitude of the laser beam projected onto the SLM. To make sure that the generated optical image would carry the information of the digital image $\boldsymbol{X}_0$, the normalized light intensity of the input optical image should be equal to the normalized pixel value of the input digital image, i.e., $\frac{|\mathbf{U}_0|^2}{{A_0}^2} = cos^2\left(\frac{\boldsymbol{\varphi}_{\mathrm{SLM}}}{2}\right) = \boldsymbol{X}_0$.

As such, a desired SLM phase pattern can be obtained by $\boldsymbol{\varphi}_{\mathrm{SLM}} = 2\arccos\left(\sqrt{\boldsymbol{X}_0}\right)$, which produces an input optical field with the expression of $\mathbf{U}_0 = A_0\sqrt{\boldsymbol{X}_0}\exp\left(i\arccos\left(\sqrt{\boldsymbol{X}_0}\right)\right)$.

Then, the SLM-generated optical image that carries the information of the input digital image incidents on the metasurface. Each meta-atom modulates the phase and amplitude of the incident light at a subwavelength scale, enabling a Hadamard product with a complex weight matrix W. The entries of the matrix are determined by the geometry of individual meta-atom. The wavefront emanating from each meta-atom acts as a source of secondary spherical waves, which interfere as they propagate and ultimately converge at the camera sensor. As a result, the optical wave prior to detection is

$$U(x,y,z) = \frac{1}{i\lambda}\iint W(x',y')U(x',y',0)\frac{ze^{ikr}}{r}dx'y' \tag{1}$$

Every point in the detection plane represents a weighted sum of all points from the input image, analogous to a fully connected neural network. Since the number of pixels in the detector is much smaller than the number of meta-atoms, the optical field at the detection plane is first downsampled through sum pooling. The detector then applies a square function to the pooled signal.

To better illustrate the relationship between the size of the meta-atoms and the camera resolution, we simplify the analysis to one-dimensional (1D) diffraction, as shown in Figure S3a. The output optical field after the metasurface is

$$\mathbf{U}_{meta} = \mathbf{U}_0 \otimes \boldsymbol{W} = \begin{pmatrix} w_1 U_0(x_1) \\ w_2 U_0(x_2) \\ \vdots \\ w_N U_0(x_N) \end{pmatrix} \tag{2}$$

w_N is a complex number characterizing the complex modulation of the *N*-th meta-atom, $U_0(x_{\mathrm{N}})$is the input optical field reaching the *N*-th meta-atom at the *x*-axis position x_{N}, and $\otimes$ denotes Hadamard product. After prorogating in space after a distance *z*, the output optical field at the detection plane before detection is

$$\mathbf{U} = \begin{pmatrix} U(x_1) \\ U(x_2) \\ \vdots \\ U(x_N) \end{pmatrix} = \boldsymbol{H}\mathbf{U}_{meta} \tag{3}$$

$\boldsymbol{H}$ is a $N \times N$ matrix characterizing 1D diffraction with entries $H_{m,n} = \frac{1}{i\lambda}\frac{e^{ik\sqrt{(x_m - x_n)^2 + z^2}}}{\sqrt{(x_m - x_n)^2 + z^2}} z$ where λ is the working wavelength, $k = 2\pi/\lambda$ is wave vector. x_m and x_n are the *x*-axis positions of *m*-th meta-atom at the image sensor plane and *n*-th meta-atom at the metasurface plane, respectively. Therefore, every point in the detection plane is the weighted sum of all modulated inputs from 41 million meta-atoms.

Since the number of pixels in the detector is much smaller than the number of meta-atoms, the optical field at the detection plane is first downsampled through sum pooling (summing neighboring elements). The detector then applies a square function to the pooled signal. Let *N* be the number of meta-atoms, *M* be the number of camera pixels, and *R* be the ratio of these two quantities (*R=N/M*) which is also the downsampling ratio. The output after the camera can then be expressed as $\widetilde{U}$ with entries:

$$\widetilde{U}_k = (\sum_{i=k\cdot R}^{(k+1)\cdot R-1} U(x_i))^2 = [(\sum_{i=k\cdot R}^{(k+1)\cdot R-1} H[i,:])\mathbf{U}_{meta}]^2 \tag{4}$$

The dimension of $\widetilde{U}$ is *M*, which equals to the pixel number (i.e., resolution) of the camera. And $\widetilde{H}_k = \sum_{i=k\cdot R}^{(k+1)\cdot R-1} H[i,:]$. Both $\mathbf{U}_{meta}$ and $\widetilde{H}_k$ are *N*-element vector.

The equation (4) implies that the system functions as a single-layer fully connected neural network (FCNN) with $N \times M$ weights, followed by *M* square functions as nonlinear activation functions. Here, *N* represents the number of meta-atoms, which is on the order of 41 million, and *M* corresponds to the camera resolution, which is on the order of 800 × 600 = 480,000. $N \times M$ weights provide very rich presentations of the input images.

Given that the metasurface, combined with optical diffraction, forms a fully connected neural network with a complex weight matrix of size $N \times M$, where *N* is the number of meta-atoms and *M* is the number of camera pixels. The operations of performing complex-valued multiplication and summation with computing such a

complex matrix are $N{\cdot}M$ and $M{\cdot}(N-1)$, respectively. One complex-valued multiplication contains 4 real multiplications and 2 real summations. One complex-valued summation contains 2 real summations. Therefore, computing such a complex matrix contains $2M{\cdot}(4N-1)$ linear operations. The number of nonlinear activation functions is M. Therefore, the total operation is 1.57×10^{14} with $N = 41$ million and $M = 480{,}000$. Considering that the metasurface perform all the linear operations in the free-space domain, the ratio of the operations performed by the metasurface and the total operations performed by the whole NN is $(1.57 \times 10^{14} - 480{,}000)/1.57 \times 10^{14} > 99.999$ %.

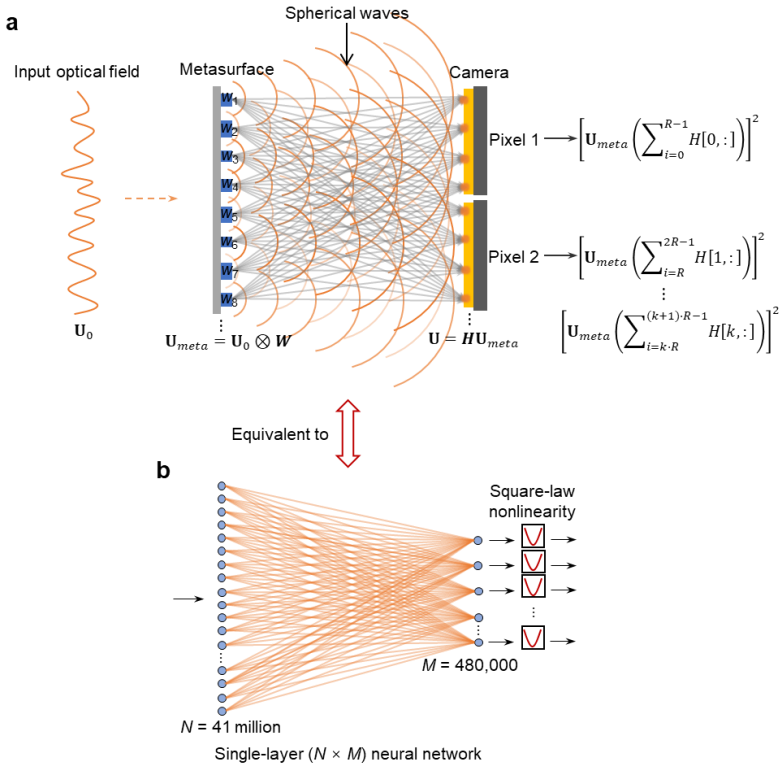


Figure S3. (a) The physical mode of the MOLM. (b) The network architecture of the MOLM.

2.Experimental setup

The experiment setup starts with a 532-nm laser diode which emits an optical beam with a diameter of 3mm and a wavelength spectrum with a full width at half maximum (FWHM) of 10 nm, as shown in Figure S4. The laser is then collimated using an optical converging lens and expanded to a diameter of 10 mm using an optical

4-*f* system. The purpose of expanding the beam diameter is to effectively cover the active area (15.4 mm × 9.6 mm) of the spatial light modulator (SLM, Meadowlark Optics E19x12-400-800-HDM8). The expanded laser beam incidents onto the spatial light modulator at an off-axis angle of ~ 12°. The SLM is sandwiched between two polarizers to encode the digital image onto the optical beam. The polarization angle of both polarizers is 45° to the horizontal direction. To ensure that the optical image can effectively cover the metasurface chip to realize the full-area modulation, the effective area of the optical image is intentionally set as 4.1 × 4.1 mm^2. Two approaches are used to realize such an area: (1) For the task images whose pixel scale is smaller than 512 × 512, the images are upsampled to a pixel scale of 512 × 512 using the nearest area interpolation method, so the effective area of the optical image is 512 × 8 μm × 512 × 8 μm = 4.1 × 4.1 mm^2. For example, the pixel scale of the COVID-19 image is upsampled to be 512 × 512 from 299 × 299. (2) For the task images whose pixel scale is larger than 512 × 512, the generated optical image is minified to be the effective area of 4.1 × 4.1 mm^2 using an optical 4-*f* system in the optical domain. For example, the Chest X-ray image with a pixel scale of 1024 × 1024 is minified double. In addition, since the effective area of the optical image is smaller than the laser beam size (10 × 10 mm^2). An aperture of ~ 5.5 × 5.5 mm^2 is used to block the stray light of the laser beam. After that, the optical image propagates through the beam splitter and is projected on the metasurface. The optical image is reflected by the metasurface and goes through another beam splitter. Following that, the optical image is focused by an optical converging lens with a focal length of 150 mm. At last, a camera (IRVI Contour IR Digital CMOS Camera) with an 8-bit depth and an activated pixel number of 600 × 800 is placed close to the focal plane of the converging lens to capture the optical image.

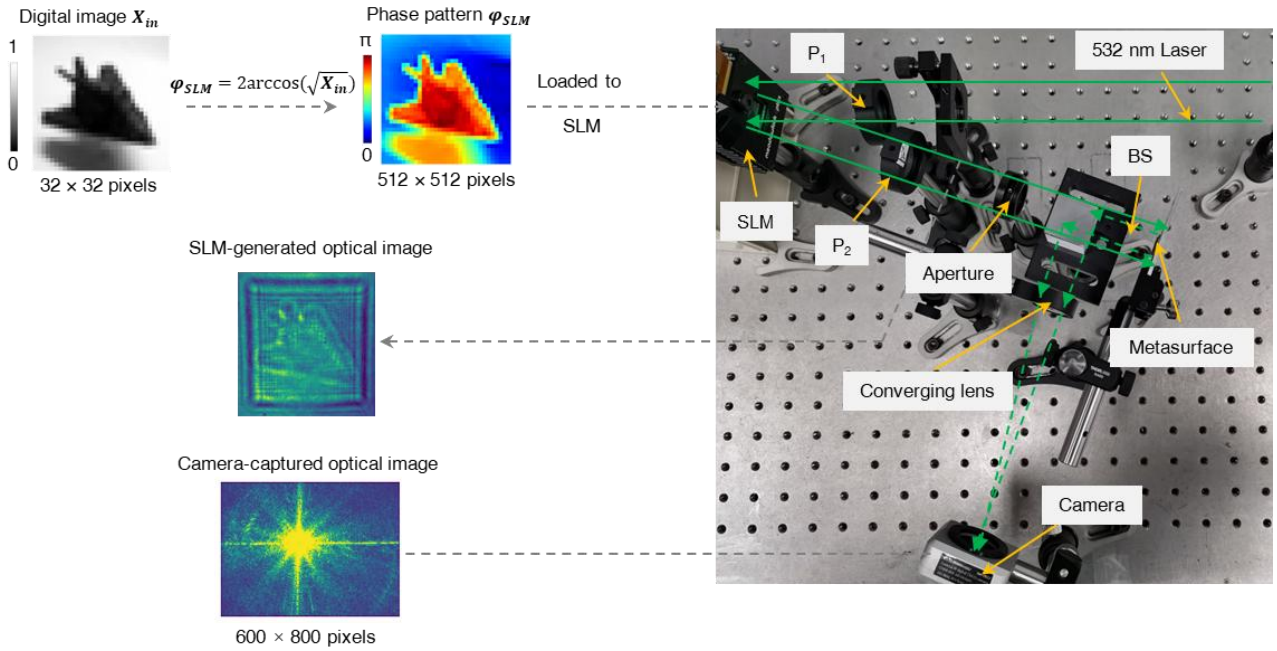


Figure S4. Experimental setup. P_1 and P_2: the linear polarizers with the same polarization directions. SLM: spatial light modulator. BS: Beam splitter. The green solid/dashed arrows represent the laser beam propagation traces before/after being reflected by the metasurface.

Note 5. Numerical simulations of a digital ELM derived from MOLM

Our system is equivalent to the original ELM formulation, the model is a simple single-hidden-layer fully connected network: the hidden-layer weights/biases are randomly assigned and fixed, and only the output layer is trained or solved by a close form. In our system, the metasurface-based optical frontend is an optical single-layer fully connected NN with random optical parameters, which do not require an end-to-end training process. A light-weight digital backend is trained using the outputs of the optical frontend to generate the prediction results.

Here we prove our hybrid is equivalent to a digital ELM in its original ELM formulation.

As shown in Figure S5a, the incident light ($\mathbf{U}_0$) carrying the encoded image incident on the optical metasurface with N meta-atoms ($M = 4.1 \times 10^7$ in the experiment). The process implements a Hadamard product ($\otimes$) between the input optical image ($\mathbf{U}_0$) and the complex-valued transmission matrix ($\boldsymbol{W}$)

determined by the metasurface. The optical image modulated by the metasurface continues to propagate in the free space and reaches the optical lens. According to the Huygens-Fresnel principle, the modulated wavefront at each meta-atom acts as secondary spherical waves and spreads out in all directions. The resulting wavefront at the optical lens is formed by the combination of all these secondary wavelets. Consequently, each meta-atom is fully connected to the optical lens, forming a $N \times N$ all-connected network with the connecting matrix ($\boldsymbol{H}_d$) determined by optical diffraction. After that, the output optical field is further processed by the optical lens that performs a Fourier transform which can be expressed by a matrix where $\boldsymbol{W}_f$ with the entry of $w_f(p,q) = e^{-i2\pi\frac{(p-1)}{m}(q-1)}$. After that the output optical field is received by the CMOS sensor with M pixels ($N = 4.8 \times 10^5$ in the experiment). Since that the pixel resolution number (N) of the camera is much less than the number of the meta-atoms (M), the camera will sum the neighboring optical field, performing an averaging pooling operation. The average pooling matrix is $P \in \mathbb{R}^{N\times M}$. The i-th row of P places $1/k$ on the indices covered by the i-th pooling window and zeros elsewhere:

$$P_{i,t} = \begin{cases} \frac{1}{k}, & t \in \{(i-1)s+1, (i-1)s+2, \dots, (i-1)s+k\} \\ 0, & \text{otherwise .} \end{cases}$$

This process leads to a fully connected $M \times N$ network. The resulted matrix is shown in Figure S5c. Here, we choose M and N as 1024 and 300 for a better visualization.

The photodetector performs a square-law operation, which serves as the nonlinear activation (neuron) in the hidden layer of the ELM. After photodetection, the digital image is first downsampled to 300 pixels using averaging pooling. This operation is equivalent to a matrix-vector multiplication. The average pooling matrix $P \in \mathbb{R}^{300\times N}$. A downsampled pixels serves as the input of a single layer of fully connected NN with 10 outputs and a sigmoid nonlinear activation function. The subsequent digital NN functions as the output layer of the ELM, shown in Figure S5b.

Consequently, the overall system is architecturally equivalent to a standard ELM, namely a single-hidden-layer fully connected network in which the hidden-layer parameters are fixed and only the output layer is trained. The key differences are that the nonlinear activation function in our hidden layer is a square function, whereas conventional digital ELMs typically employ sigmoidal activation functions, and that the fully connected weight matrix in our system is complex-valued and governed by the diffraction operator, while in standard ELMs the weights are arbitrary real-valued random variables.

As shown in Figure S5d, both the digital ELM and our system exhibit improved performance as the hidden-layer width increases. When the network width is increased to 41 million neurons, the MNIST classification accuracy of the digital ELM reaches approximately 99% using diffraction-limited complex-valued random weights, which is close to the experimental accuracy of 99.2% reported in our manuscript. In the simulations, the simulated ELM is illustrated in Figure S5b. The number of output nodes (N) of is fixed to be 300. For the second layer (output layer), the input size of the second layer is 300 and the output size is 10. The M and N are chosen due to the computational limits of our simulation platform. In the actual experiment, M = 41 million, and N = 480,000.

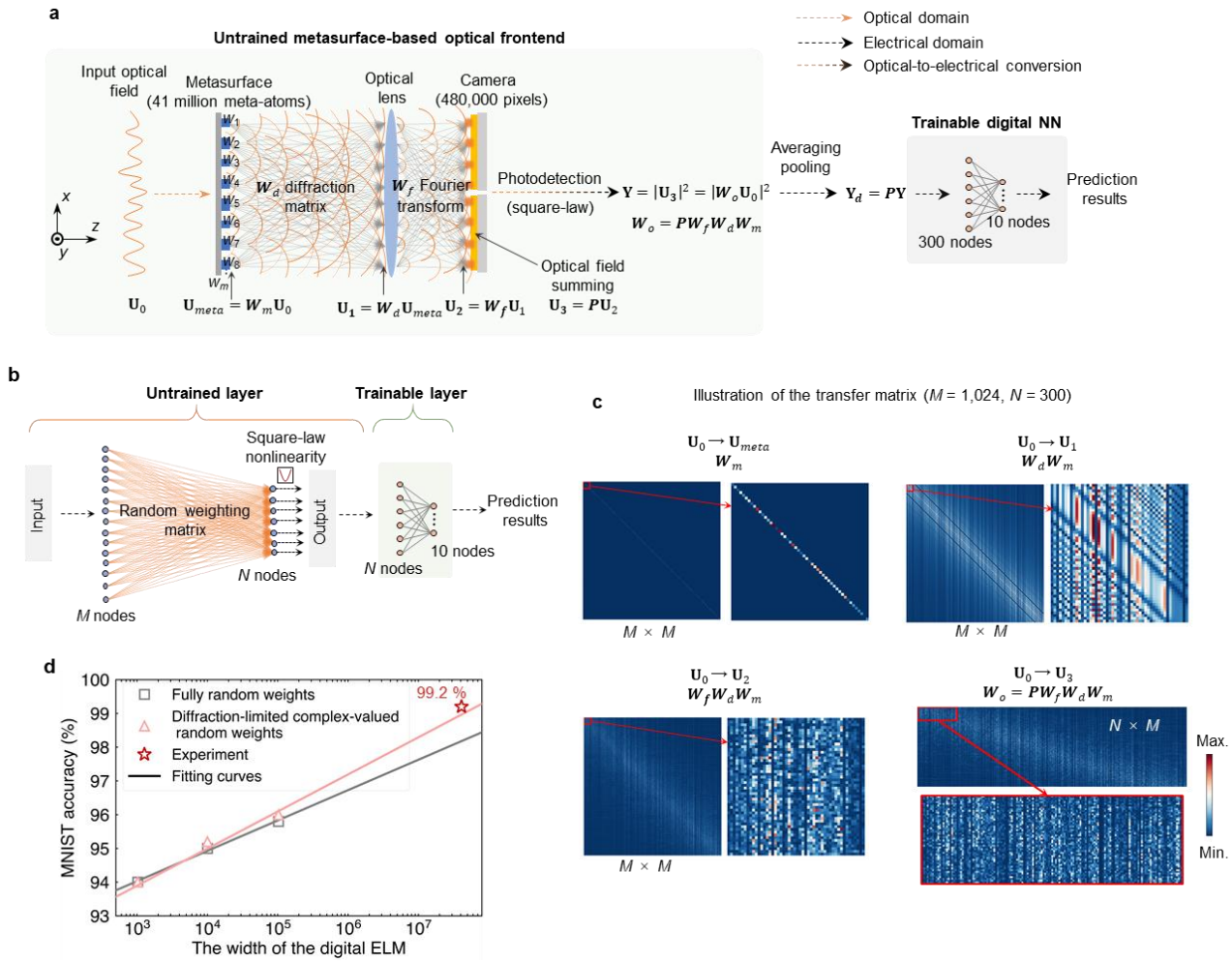


Figure S5. (a) The physical process of our metasurface-based computing system. (b) The network architecture of the NN derived from our system. (c) Illustration of the modulus of various complex-valued transfer matrices of our system. (d) The simulated MNIST accuracies versus the width of the ELM.

Note 6. Numerical simulations of the MOLM based on the MNIST dataset

We utilize the network model, as introduced in **Note 4**, to simulate the performance of the MOLM on the MNIST dataset. The complex-valued optical parameter of the meta-atom consists of phase and amplitude, which are determined by the diameter of the meta-atom, as introduced in **Note 1**. In the simulation, the diameter of the meta-atom in the optical frontend and the digital weights in the backend are trainable to minimize the cross-entropy loss. Figure S6 illustrates the phase modulation coefficients for the trained and untrained metasurfaces and corresponding transformed images from the metasurface-based system. The transformed

image of the metasurface-based optical frontend is downsampled to 15 × 20 from the original size of 600 × 800. The optimization algorithm is Stochastic Gradient Descent (SGD). 9,000 images and 1,000 images are used for training and testing, respectively. The codes are run on Python 3.9.0 and TensorFlow 2.8.0, based on a GPU RTX 4090.

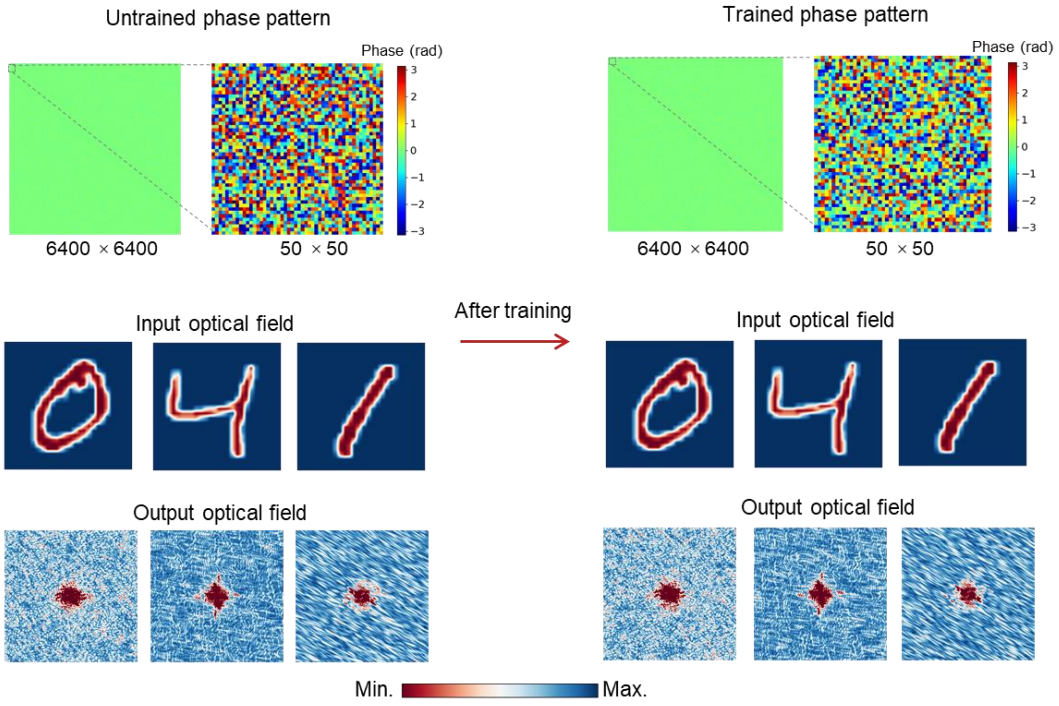


Figure S6. The phase modulation coefficients for the trained and untrained metasurfaces, along with the corresponding optical fields for the MNIST dataset.

Note 7. Quantitative phase imaging for characterizing the complex-valued modulation matrix of the metasurface with a sub-wavelength resolution

To show our metasurface indeed provides a complex-valued modulation matrix for the incident light, Quantitative Phase Imaging (QPI) [2] is used to measure the optical phase response of several local regions of the metasurface, as shown in Figure S7a. A single-mode fiber coupled 532 nm laser is used to illuminate the sample under normal incidence. A 50× objective lens (TU Plan Fluor Epi P, NA 0.8, Nikon) is used for sample illumination and scattered light collection, which provides a diffraction-limited spatial resolution of 665 nm.

By magnifying the reflected optical field from the metasurface by 200× to the camera (Q-2A750/CXP, Adimec), the system achieves a high imaging pixel resolution of 60 nm and a field of view of 36 × 36 μm^2. Figure S7b shows 4 corners and 1 center region of the metasurface. The imaging field of view is small but can be increased by using higher resolution camera. The phase within the metasurface changes from – 2 to 2 rads while the region out of the metasurface has a uniform phase. This indicates effective phase modulation with sub-wavelength resolution is provided by the metasurface. The amplitude modulation pattern is measured from the green channel of the optical microscope image, as shown in Figure S7d. The measured amplitude modulation pattern also exhibits spatial changes ranging from 0.5 to 1. Figures S7e–f gives the statistical histogram of the phase and amplitude values of the measured result, which can be fitted well by a Gaussian distribution. The statistical results indicate that the modulation matrix of the metasurface follows an optimized Gaussian distribution in both phase and amplitude domains.

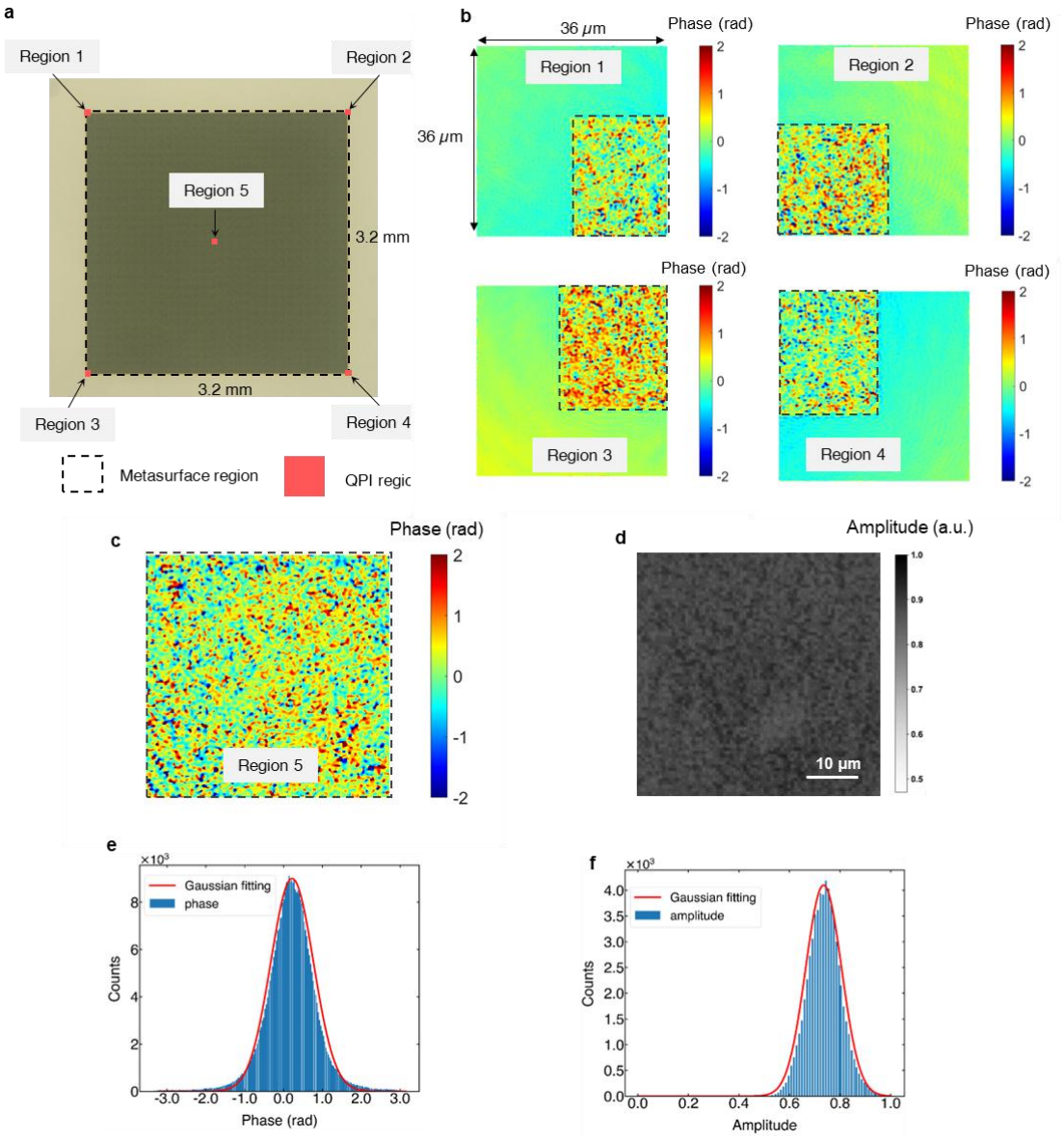


Figure S7. Reflective quantitative phase imaging (QPI) for the characterization of the complex-valued modulation of the metasurface. (a) The optical microscope image of the metasurface chip. The black dashed box represents the boundary of the effective area of the metasurface. The red solid boxes represent the view fields of QPI. (b) The phase modulation pattern of the various metasurface regions extracted from QPI. (c) The phase modulation pattern of a local region of the metasurface. (d) The amplitude modulation pattern of a local region of the metasurface. (e) The statistic histogram of the phase values of the QPI measured result. (f) The statistic histogram of the amplitude values of the measured result.

Note 8. Characterization of reflection efficiency of metasurface

An optical power meter (Thorlabs S120VC) is used to measure the power of the reflected light by the metasurface, as shown in Figure S8a. A pure-white digital image is encoded to the spatial light modulator to generate a uniform laser beam. The optical powers before and after being reflected by the metasurface are meausred, respectively. The reflection efficiency is the ratio of these two measured powers. As shown in Figure S8b, the reflection efficiency ranges from 20.2 % to 25.0 % at various incident angles. The corresponding insertion losses range from 7.0 dB to 6.0 dB. Here, the incident angle is with respect to the normal direction. The insertion loss can be reduced by using low-loss materials, for example, sapphire and titanium dioxide.

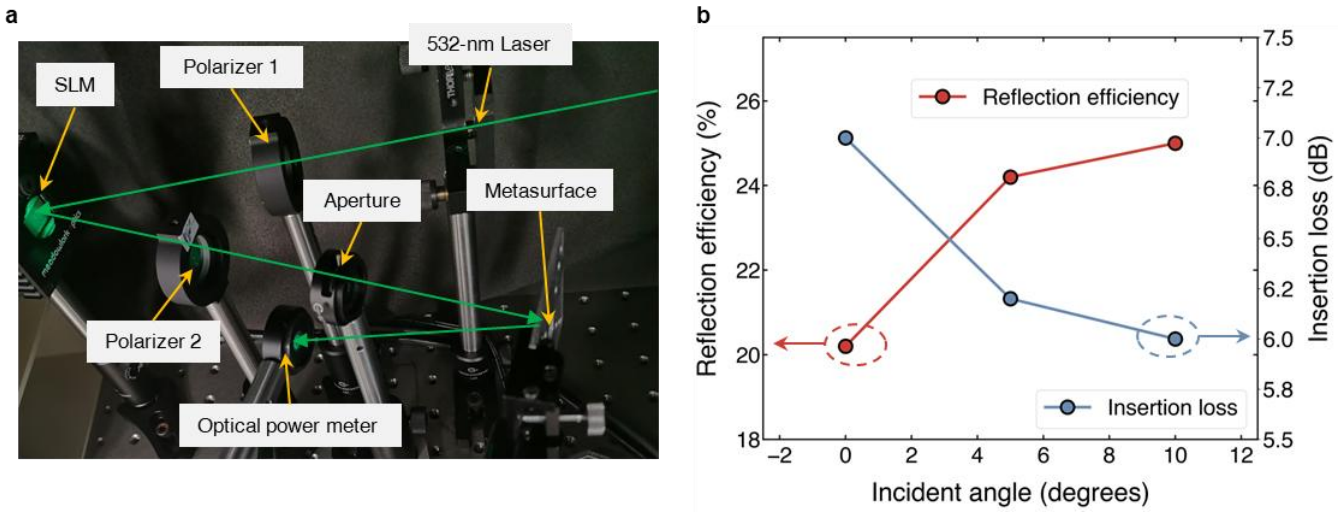


Figure S8. The measured reflection efficiencies and insertion losses of the metasurface. (a) The experimental setup for the measurement of the reflection efficiency of the metasurface. SLM: spatial light modulator. The green solid arrows represent the laser beam propagation traces. (b) The measured reflection efficiencies and insertion losses of the metasurface versus the incident angles.

Note 9. Evaluation of impact of optical parameter number on image classification accuracy.

To validate that the superior performance of the MOLM is a result of the availability of a large number of optical parameters uniquely provided by metasurface, we first experimentally tested the image classification accuracy when optical parameter numbers (i.e., nanodisk numbers) vary from 4.2 million to 41.0 million. The

total pillar number of the metasurface is adjusted by changing the area of the metasurface that the optical image incident on. For example, COVID-19 dataset image is downsampled or upsampled to a pixel scale of 128 ×128, 256 × 256, and 512 × 512 using the nearest area interpolation method. Those images are then loaded into the SLM. As such, the SLM-generated optical images can illuminate three different areas of the metasurface in the experiment, i.e., 1.024 × 1.024 mm^2, 2.048 × 2.048 mm^2, and 3.20 × 3.20 mm^2. The optical parameter area density of the metasurface is 4 million per mm^2, so we can achieve the various optical parameter numbers which are 4.2 million, 16.8 million, and 41 million, respectively.

Note 10. Characterization of the ability of fast recovering from perturbations

Using the human action recognition task, we experimentally investigated the robustness of the MOLM under a physical perturbation. The preparation of the experimental dataset is the same as introduced in **Note 16**. The physical perturbation is realized by applying an axial misalignment of ~ 10 μm to the metasurface in the experimental setup introduced in **Note 4**. Before applying the perturbation, the experimental accuracy reaches 99.1 %. After applying the perturbation, we repeat the above experiment to obtain the new metasurface-processed images and the accuracy drops to 16.7 %, as shown in Figure S9. The significant reduction in accuracy is the result of the change in the optical transfer function of the MOLM, arising from the axial misalignment of the metasurface. To recover the classification accuracy, the system is retrained using 1,920 images. After 45 training epochs using a GPU of NVIDIA GeForce RTX 3090, the accuracy is recovered to 96.3 % within a short period of 234 ms, tested from 1500 testing images. The experimental result indicates that our MOLM can quickly recover from the physical perturbations, which enables the deployment of the MOLM in real-time fast-decision applications.

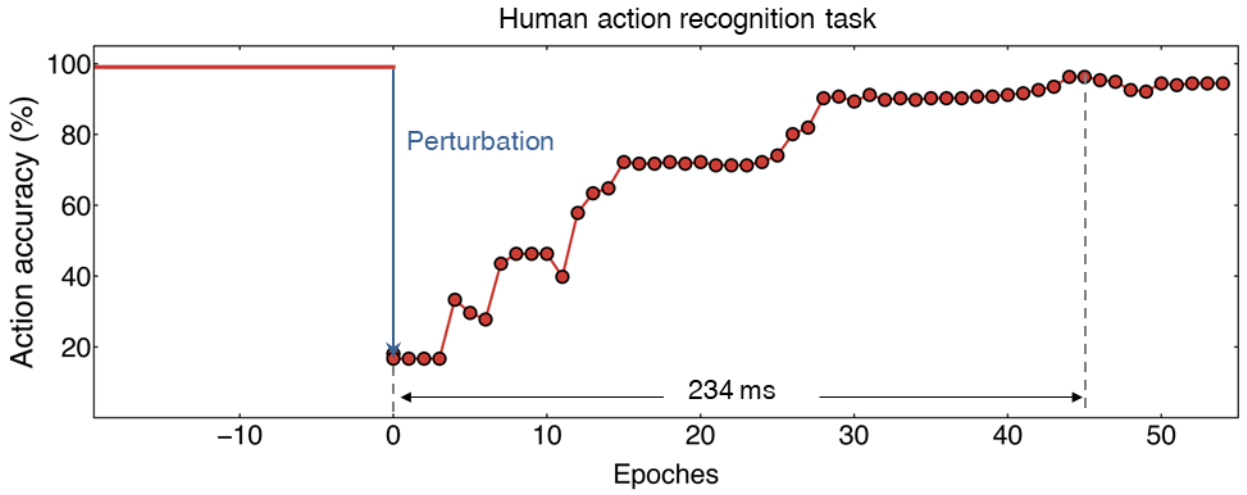


Figure S9. Experimental classification accuracies of the human action recognition task versus retraining epochs after applying the perturbation.

Note 11. Characterization of the MOLM operating under different illumination conditions

We experimentally validate the ability of our metasurface to operate even under low-illumination conditions. The ability to work under low illumination condition is very important for tasks involving dark or underexposed environments such as shooting high-speed moving objects. To emulate different illumination conditions, in the experiment, the light power is attenuated using the optical attenuators that are placed in front of the laser source. We use the dataset image from the human action recognition to evaluate this capability. The images are loaded to SLM to generate the corresponding optical image, which are then filtered by a hole-shape aperture with a diameter of ~ 5.5 mm before being detected by the optical power meter, as shown in Figure S10a. By varying the optical attenuators, a series of optical powers are obtained, as shown in Figure S10b. In the table of Figure S8b, the average optical energy E_i received by one meta-atom that implements a single optical parameter is calculated using the following equation:

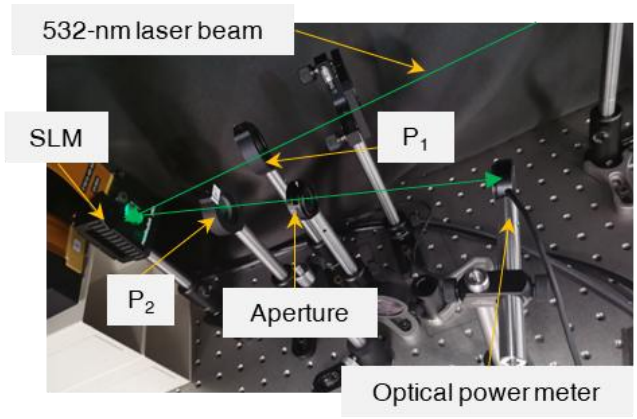
$$E_i = \frac{p_{in} \times t_e}{N_{neuron}}$$

where p_{in} is the total power of the input laser beam received by the metasurface, t_e is the exposure time of the camera, and $N_{neuron}$is the number of total optical parameters. We take t_e = 34 μs, and N_{neuron} = 41 million

into the above equation to obtain the average optical energy received by one meta-atom. One meta-atom implements one complex-valued multiplication, so the detected photon number per complex-valued multiplication N_p can be expressed as follows:

$$N_P = \frac{E_i}{E_p}\,\eta$$

E_p is the energy of single photon at a wavelength of 532 nm and η is the reflection efficiency of the metasurface under normal incidence, i.e., 20.5 % as the detected light is the light reflected by the metasurface. The video classification accuracy is evaluated at different input powers as shown in Figure S10c. The accuracy can reach 95.0% when the received optical power by the camera is only 0.082 μW, corresponding to a detected photon number of 0.078 required for one complex-valued multiplication. The result shows that our MOLM can operate effectively with low energy consumption and in low light conditions.



b

Input power p_{in} (μW)	Average optical energy received by one neuron E_i (10^{-19} J)	Detected photons per complex-valued multiplication N_d
9.0	74.7	4.10
0.91	7.55	0.41
0.17	1.411	0.078
0.086	0.714	0.039
0.014	0.116	0.0064

c

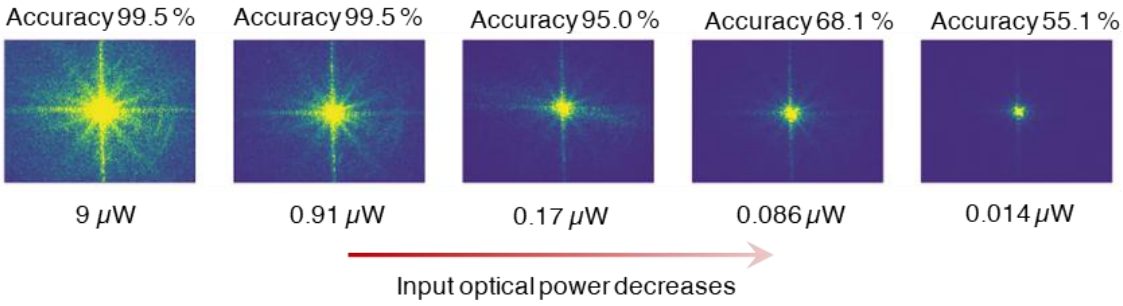


Figure S10. (a) The experimental setup for the measurement of the light power of the SLM-generated optical image. SLM: spatial light modulator, P_1 and P_2: linear polarizer. (b) The table of the summary of the measured light power and related parameters. (c) The camera-captured images of the human action recognition task under various input optical powers.

Note 12. Characterization of the MOLM operating under a broadband light source

To demonstrate the ability of our system to work under a broadband light source, we have also evaluated the MNIST accuracy of our system under a broadband light source with different spectral widths ranging from 20 nm to 100 nm. The LED source has a broadband continuous spectrum covering the entire visible regime, as shown in Figure S11a. For the 20-nm spectral width, the LED light contains all wavelengths ranging from 520 nm to 540 nm. For the 50-nm spectral width, the LED light contains all wavelengths ranging from 500 nm to 550 nm. For the 100-nm spectral width, the LED light contains all wavelengths ranging from 500 nm to 600 nm. For comparison, we used a monochromatic coherent laser source with a wavelength of 532 nm.

Figure S11b shows that the experimental MNIST accuracy can reach over 98.3 % under the incoherent LED source for the spectral width of 20 nm and 50 nm, with a slight degradation of less than 1 % compared with the 532-nm coherent laser source. By further expanding the spectral width to 100 nm, the MNIST accuracy can still preserve 97.3 %. The reason for the degradation is that the incoherent ELD light can only realize linear computations in the intensity domain, while the coherent laser can realize perform linear computations in the complex-valued domain. This leads to a reduction of model complexity in an LED-based computing system. Nevertheless, the MNIST accuracy under the broadband incoherent light can still outperform that of most current photonic ELMs that adopt the untrained optical front, all of which work under the coherent laser source. The above experimental results indicate that our system can work under a broadband incoherent light source, which is highly required in practical applications. In the experiments, the accuracy is obtained by averaging the accuracies from 10 different splitting configurations of the dataset.

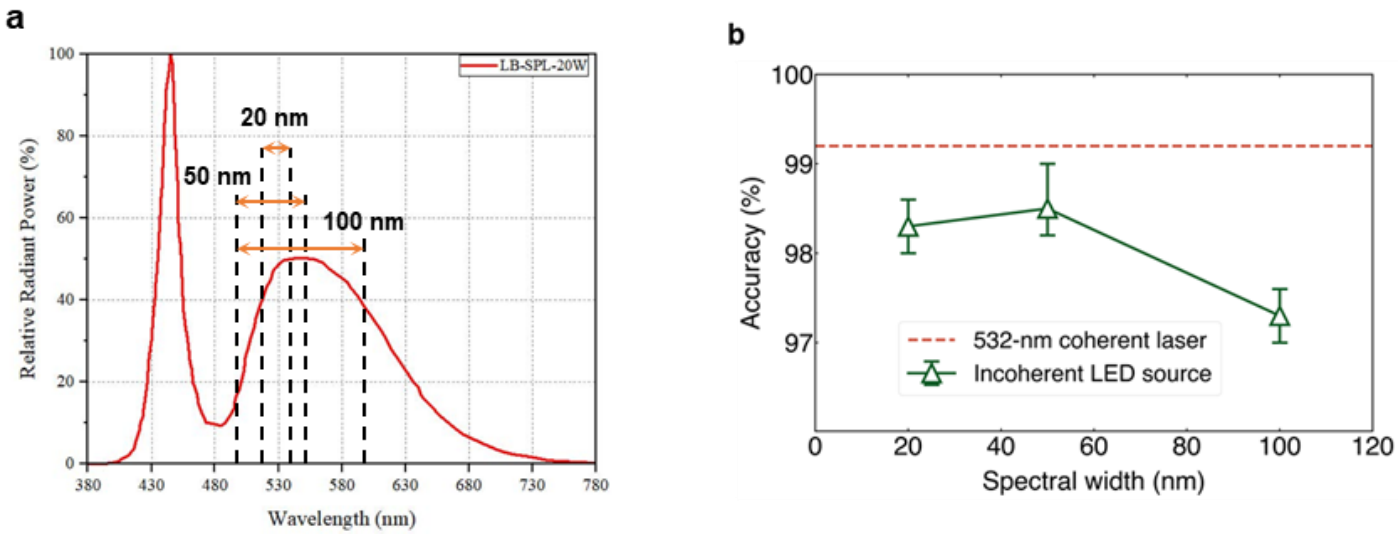


Figure S11. (a) The emission spectrum of the LED source used in the experiment. (b) The MNIST accuracies measured under an incoherent LED source with various spectral widths.

Note 13. Performance comparison with random scatterers

We performed a comparative experiment between a metasurface and a random scattering medium. The random scattering medium is fabricated by frosting the surface of a transparent glass substrate, as shown in Figure S12a. The rough frosted surface consists of many irregular particles that can disperse the incident light to form a transmission pattern with random speckles. By controlling the fineness of the abrasive, the size of the rough particles can be varied to provide different levels of light diffusion. We used a random scatterer medium with approximately 1,600 meshes in a 25.4-mm region, corresponding to the size of the rough particles that varies near 10 μm (i.e., the finest feature size in commercially available products). Under otherwise identical experimental conditions, the optical metasurface achieves a classification accuracy of 99.2% for MNIST digit classification, whereas the random scattering medium yields an accuracy of 92.0%, as shown in Figure S12b.

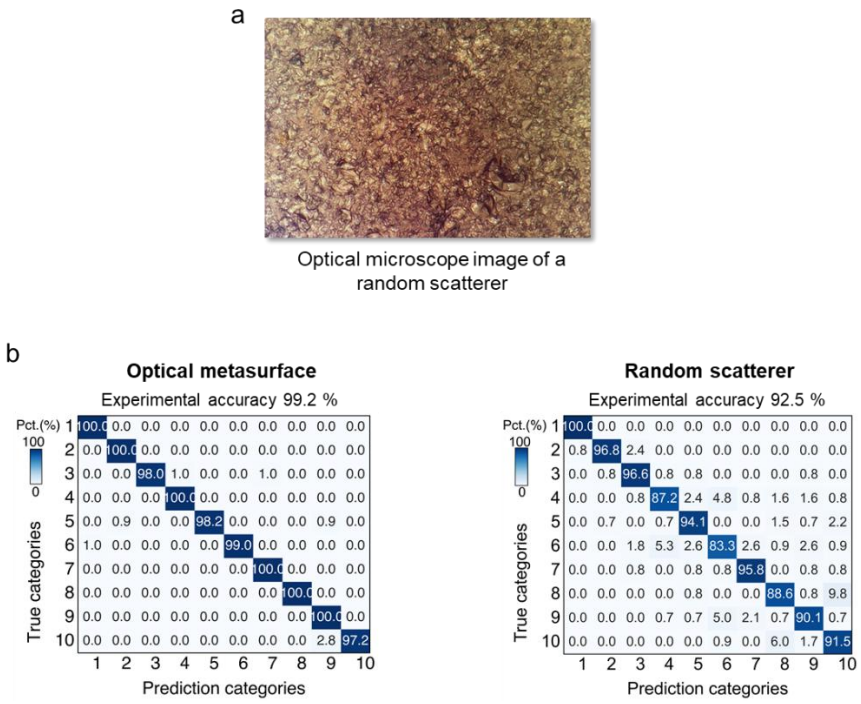
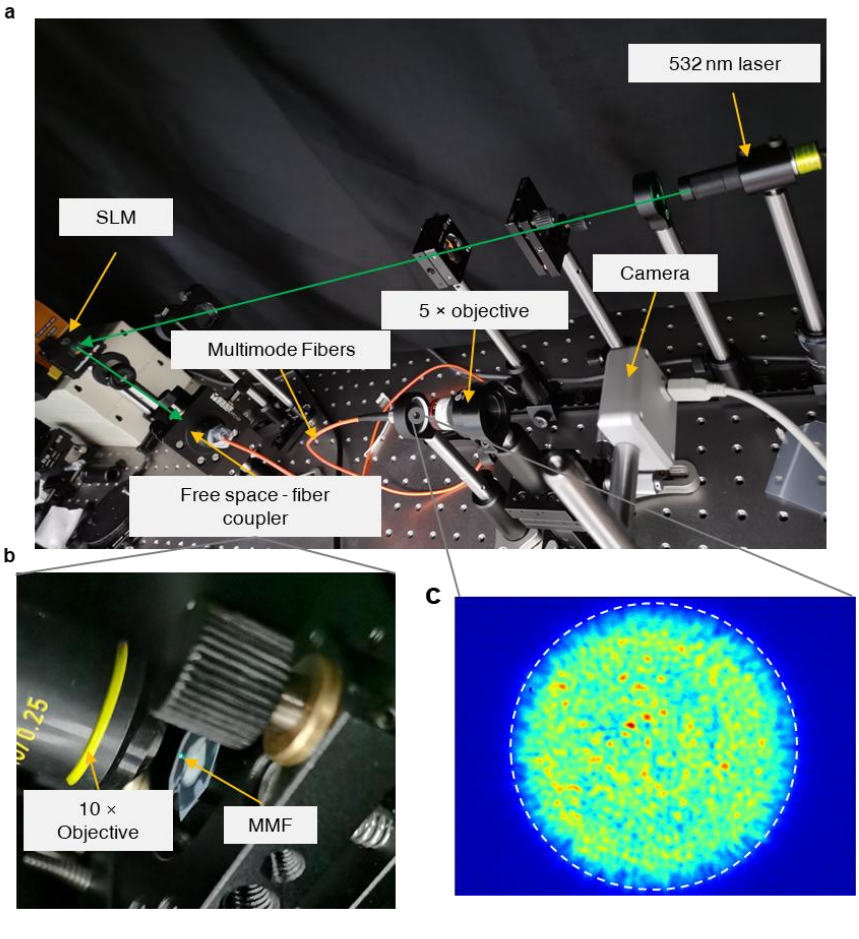


Figure S12. The experimental confusion matrix of MNIST accuracy for our optical metasurface (left-hand-side graph) and random scatterer (right-hand-side graph).

Note 14. Comparison with multimode fiber-based optical neural network

The optical multimode fiber has been used to construct an optical neural network for image classification [3–4]. The coupling between various modes as they propagate in the multimode fiber is used to construct an optical neural network whose weights are random and fixed. However, the number of modes (i.e., optical parameters) in a multimode fiber typically range from a few thousand to tens of thousands, which is greatly less than the parameter number provided by our metasurface. Consequently, our MOLM is expected to outperform the optical neural network constructed by the multimode fibers. We built up a multimode fiber-based optical neural network to compare to our MOLM, as shown in Figure S13a. The experimental setup to generate optical images is the same as that used in the MOLM. The core diameter of the multimode fiber (Thorlabs FG105LVA) used in the experiment is 105 μm. A free-space fiber coupler is used to shrink the optical image and couple it into the multimode fiber, as shown in Figure S13b. At the output of multimode fiber, the optical field is magnified

by a 10× objective and then captured by a camera. The camera-captured image is shown in Figure S13c. The camera-captured images are used to train a single-layered digital neural network and generate prediction results. Here, we evaluate two datasets: intracranial hemorrhage detection based on CT images and breast cancer detection based on pathology images. Figure S13d shows the accuracy comparison between the multimode fiber and the MOLM. The digital weight number of the single-layered neural network is kept the same in two systems for fair comparison. The accuracy of the MOLM is significantly higher than that of the multimode fiber-based ONN for both tasks.



d

Dataset	Multimode fiber	Our metasurface
RSNA Intracranial Hemorrhage Detection based on CT Images	70 %	**96.2 %**
Breast Cancer Detection based on Pathology Images	83.8 %	**95.1 %**

Figure S13. The experimental setup and results of the multimode-fiber-based optical neural network. (a) The overall experimental setup. SLM: spatial light modulator. The green solid arrows represent the laser beam propagation traces. (b)The zoom-in image of the free-space fiber coupler. (c)The camera-captured output image from the multimode fiber end facet. The white dashed line represents the boundary of the fiber core. (d) Performance comparison between the multimode fiber-based ONN and the MOLM.

Note 15. Training neural network at the digital backend

After being processed by the MOLM, the final prediction results of various machine vision tasks in the experiments are produced by a digital neural network. As shown in Figure S14a, the optical images are processed by the MOLM and then captured by the camera with a pixel size of 600 × 800. Then the captured images are downsampled by averaging the received power in the neighboring sensor array, as shown in Figure S14b, where w and d are the pixel scale of the averaging region and the averaging stride, respectively. After that, the down-sampled images serve as the new dataset. The dataset is then randomly split into the training dataset and testing dataset. The split ratio will be given in the respective tasks in the following sections.

The digital neural network consists of an input layer for receiving the input image and an output layer for generating prediction results, and the two layers are fully connected, as shown in Figure S14c. Biases are added to the output before the activation function of *sigmoid*. The mathematical expression of the single-layered digital neural network is $\boldsymbol{y} = sigmoid\ (\boldsymbol{b_0} + \boldsymbol{w_0 x})$, where b_0, w_0, x, and y are the biases, the digital weights, the input image in the form of vector and the prediction result, respectively. Logistic regression is used to optimize the biases and digital weights. Finally, the testing dataset is used to verify the performance of the overall system. In the manuscript, the digital weight number is defined as the weight number of this single-layered digital NN.

The details about the parameters used in the single-layered digital neural network will be given in the respective tasks in the following sections.

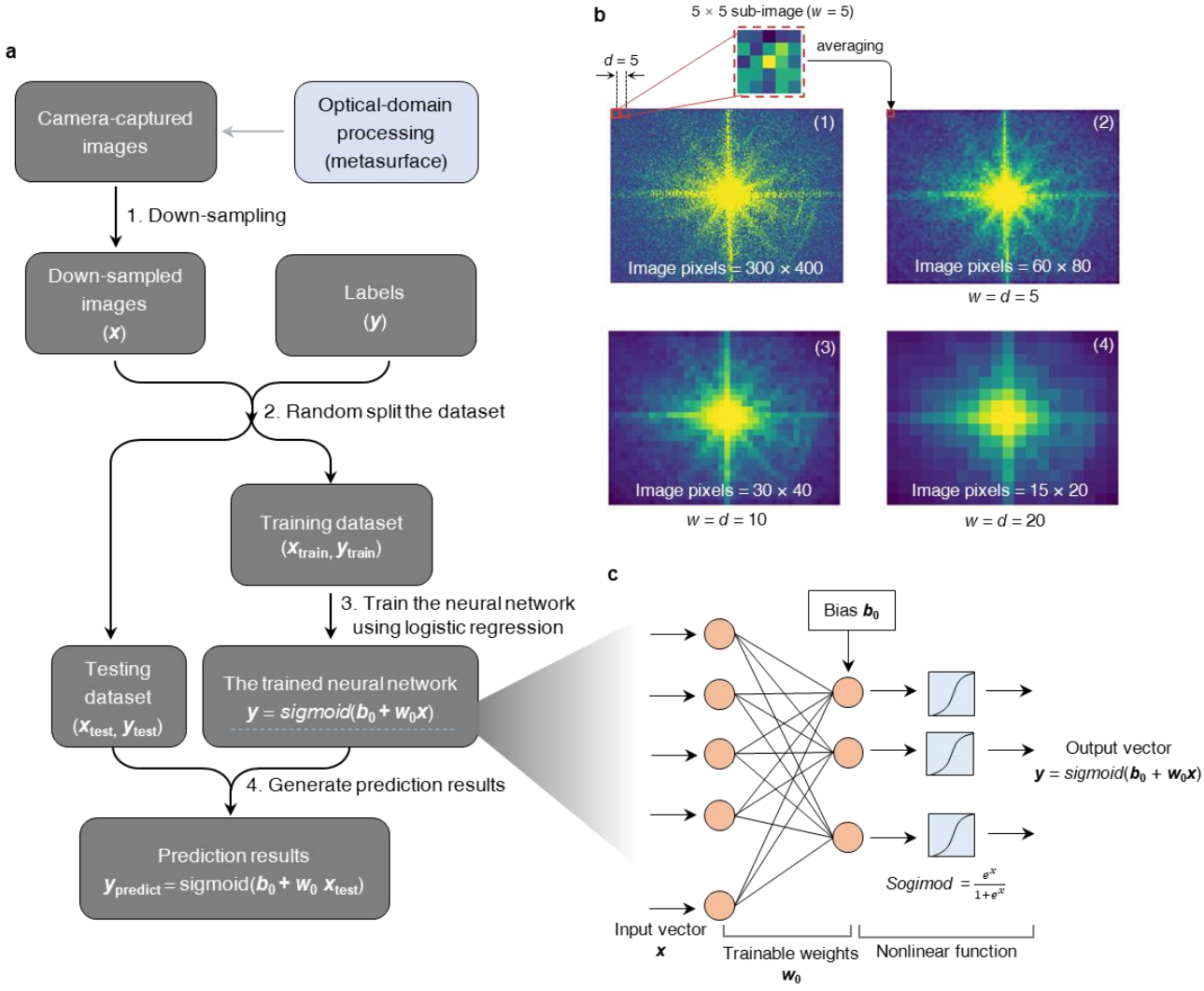


Figure S14. The details about the training of the single-layered digital neural network. (a) The block diagram of the training of the digital neural network. (b) The illustration of down-sampling the experimentally captured images. (c) The architecture of the single-layered digital neural network.

Note 16. Benchmarking task based on the MNIST dataset

(1) MNIST: The full MNIST dataset consists of 70,000 28 × 28 grayscale 8-bit images in 10 classes. The dataset used in the experiment can be found at https://www.kaggle.com/datasets/hojjatk/mnist-dataset. The MNIST images are labeled with one of the following 10 categories: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. We extract 10,000 images from the full dataset to create a new dataset for the experiments, and every category has 1,000 images, as shown in Figure S15a. The dataset is randomly split into the training dataset and testing dataset with a ratio of 90% : 10%, i.e., 9,000 images for training and 1,000 images for testing. As shown in Figure S15b, the category distributions of the training and testing dataset are given. In the experiment, the original images are upsampled to a pixel scale of 512 × 512 using the nearest area interpolation method and then encoded into the SLM to generate the optical images. The purpose of resizing the image is to ensure that the effective area of the SLM-generated optical image is large enough to fully cover the metasurface chip. Followed by that, the camera is used to capture the images processed by the MOLM. The captured images are downsampled to a pixel scale of 15 × 20 using the method introduced in **Note 15**. Figure S15c lists the parameters for downsampling camera-captured images. At last, the downsampled images are used to train a single-layered digital neural network whose weight number is 15 × 20 × 10 = 3,000. Logistic regression is used to optimize the digital weights of the neural network by iterations. As shown in Figures S15d − e, the training accuracy converges to 99.5 % after 100 iterations and the testing accuracy is 99.2 %.

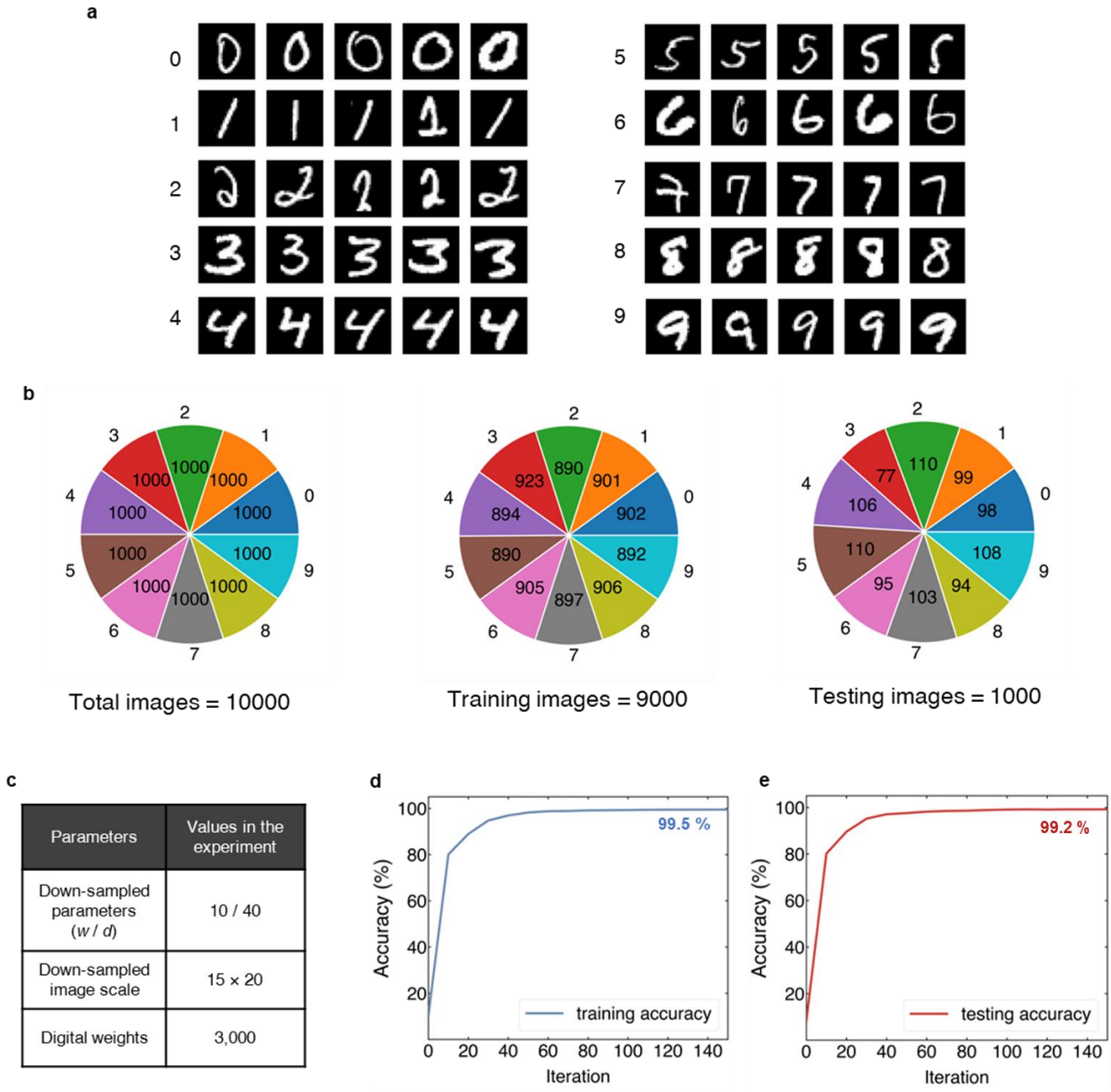


Figure S15. (a) The illustration of the MNIST dataset. (b) The category distribution of the full dataset, training dataset, and testing dataset. The number in the pie chart indicates the image number of every category. (c) The summary table of the parameters in the training of the single-layered digital neural network. (d) The training accuracies versus iterations. (e) The testing accuracies versus iterations.

(2) CIFAR-10: The full CIFAR-10 dataset consists of 60000 32 × 32 colorful images in 10 classes. The dataset used in the experiment can be found at https://www.cs.toronto.edu/~kriz/cifar.html. In the experiment, we convert the colorful images into grayscale to encode them to the SLM. As shown in Figure S16a, the CIFAR-10 grayscale images are labeled with one of the following 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. 2000 images are extracted from the full dataset to create a new dataset for the experiment, as shown in Figure S12b, every category has 200 images. The dataset is randomly split into the training dataset and testing dataset with a ratio of 75 % : 25%, i.e., 1500 images for training and 500 images for testing. The category distributions of the training dataset and testing dataset are shown in Figure S16b. In the experiment, the original images are upsampled to a pixel scale of 512 × 512 using the nearest area interpolation method and then encoded into the SLM to generate the optical images. The purpose of resizing the image is to ensure that the effective area of the SLM-generated optical image is large enough to fully cover the metasurface chip. The captured images are downsampled to a pixel scale of 24 × 31 using the method introduced in Section 3.1. Figure S16c lists the parameters for downsampling camera-captured images. At last, the downsampled images are used to train a single-layered digital neural network whose weight number is 744 × 10 = 7440. Logistic regression is used to optimize the digital weights of the neural network. As shown in Figures S16d–e, the training accuracy converges to 99.5% after 100 iterations and the testing accuracy is 91.6 %.

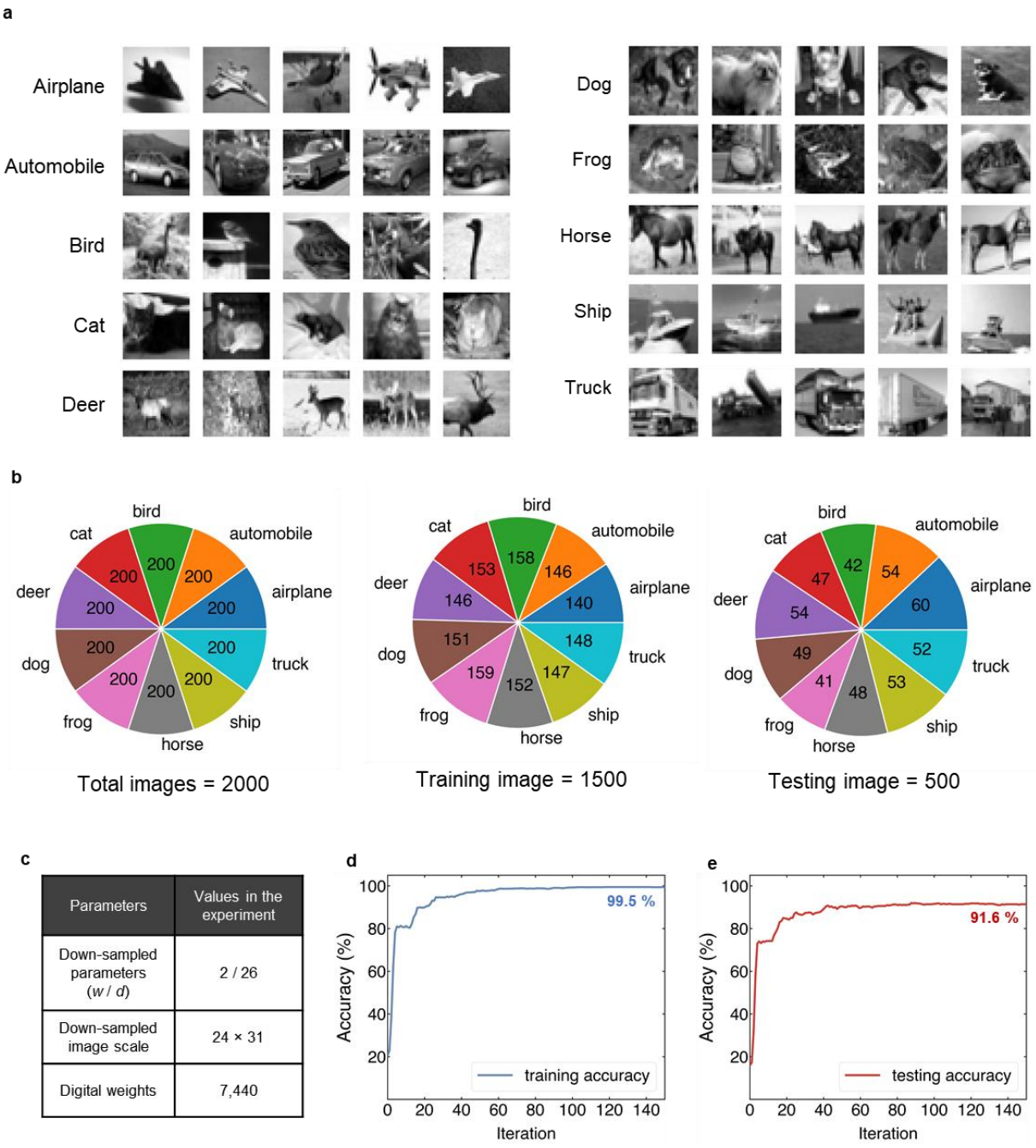


Figure S16. (a) The illustration of the grayscale CIFAR-10 dataset. (b) The category distribution of the full dataset, training dataset, and testing dataset. The number in the pie chart indicates the image number of every category. (c) The summary table of the parameters in the training of the single-layered digital neural network. (d) The training accuracies versus iterations. (e) The testing accuracy versus iterations.

Note 17. COVID-19 detection based on Chest X-ray images

The full COVID-19 Radiography dataset consists of 299 × 299 chest X-ray (CXR) images in 4 classes, which are 3616 COVID-19 positive, 10192 normal, 6012 lung opacity, and 1345 viral pneumonia images, respectively. The dataset for the experiment can be found at https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database. 1000 COVID-19 positive images and 1000 normal images are extracted from the full dataset to create a new dataset for the experiment, as illustrated in Figure S17a. The dataset is randomly split into the training dataset and testing dataset with a ratio of 75 % : 25%, i.e., 1500 images for training and 500 images for testing. As shown in Figure S17b, the category distributions of the training and testing dataset are given. In the experiment, the original images are upsampled to a pixel scale of 512 × 512 using the nearest area interpolation method and then encoded into the SLM to generate the optical images. The purpose of resizing the image is to ensure that the effective area of the SLM-generated optical image is large enough to fully cover the metasurface chip. Followed by that, the camera is used to capture the images processed by the MOLM. The captured images are downsampled to a pixel scale of 12 × 16 using the method introduced in **Note 15**. Figure S17c lists the parameters for downsampling camera-captured images. At last, the downsampled images are used to train a single-layered digital neural network whose weight number is 192 × 1 = 192. Logistic regression is used to optimize the digital weights of the neural network by iterations. As shown in Figures S17d − e, the training accuracy converges to 98.2 % after 100 iterations and the testing accuracy is 98.0 %.

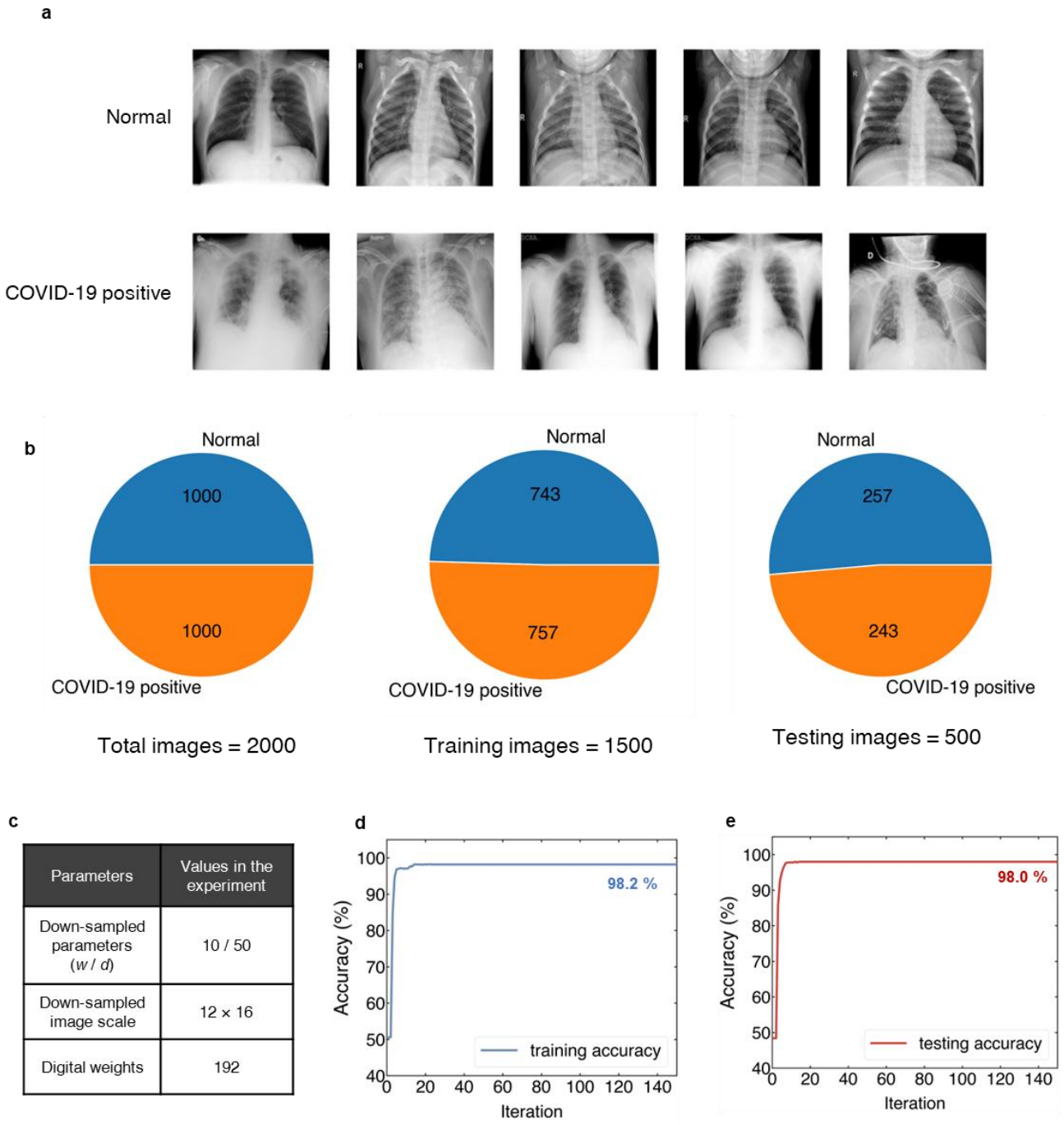


Figure S17. (a) The illustration of the COVID-19 detection dataset. (b) The category distribution of the full dataset, training dataset, and testing dataset. The number in the pie chart indicates the image number of every category. (c) The summary table of the parameters in the training of the single-layered digital neural network. (d) The training accuracies versus iterations. (e) The testing accuracies versus iterations.

Note 18. Intracranial hemorrhage (ICH) detection based on Brain Computed Tomography (CT) images

The full RSNA Intracranial Hemorrhage (ICH) Detection dataset consists of more than 25,000 512 × 512 brain computed tomography (CT) images in two classes of intracranial hemorrhage and normal. The intracranial hemorrhage CT images are divided into 5 sub-types according to the bleeding locations in the brain. The dataset used in the experiment can be found at https://www.kaggle.com/c/rsna-intracranial-hemorrhage detection/overview. 932 intracranial hemorrhage and 1,068 normal CT images are extracted from the full dataset to create a new dataset for the experiment, as shown in Figure S18a. The intracranial hemorrhage CT images for the experiment include all 5 sub-types. The dataset is randomly split into the training dataset and testing dataset with a ratio of 75 % : 25%, i.e., 1,500 images for training and 500 images for testing. The category distributions of the training dataset and testing dataset are shown in Figure S18b. In the experiment, the original images are directly encoded into the SLM without upsampling to generate the optical images. After that, the camera is used to capture the images processed by the MOLM. The captured images are downsampled to a pixel scale of 12 × 16 using the method introduced in **Note 15**. Figure S18c lists the parameters for downsampling camera-captured images. At last, the downsampled images are used to train a single-layered digital neural network whose weight number is 192 × 1= 192. Logistic regression is used to optimize the digital weights of the neural network by iterations. As shown in Figures S18d − e, the training accuracy converges to 99.1 % after 100 iterations and the testing accuracy is 98.8%.

In the task of localizing the bleeding region inside the brain CT images, 530 images are extracted from the full dataset to form a new dataset, which is randomly split into 481 images for training and 49 for testing. These images are labeled with a 1× 4 vector, which consists of the horizontal position, vertical position, width, and height of the box representing the bleeding region. The raw brain CT images are processed by the MOLM and downsampled to 15 × 20 pixels using a parameter of $w = d = 40$ as introduced in **Note 15**. The downsampled images are used to train a digital neural network, which consists of one input layer for receiving the input image, three hidden layers, and one output layer for generating a 1× 4 vector. The nodes of each hidden layer

are 64, 128, and 512, respectively, so the total parameters of the digital neural network are 97,732. The nonlinear connection between hidden layers is a *ReLu* function and the nonlinear function of the output layer is a *sigmoid* function. The loss function is the mean absolute error between the coordinates of the predicted box and the ground truth.

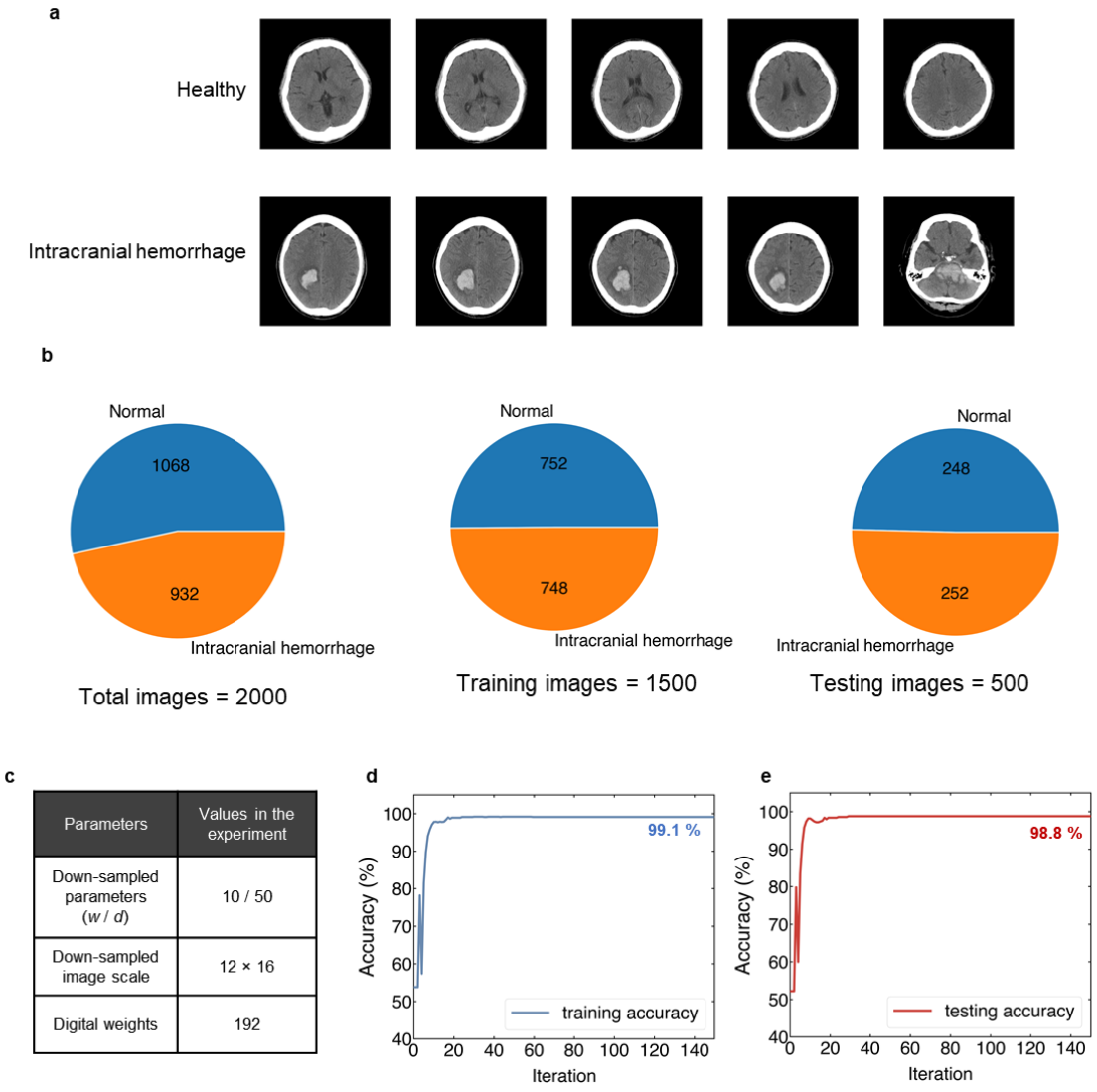


Figure S18. (a) The illustration of the intracranial hemorrhage detection dataset. (b) The category distribution of the full dataset, training dataset, and testing dataset. The number in the pie chart indicates the image number of every category. (c) The summary table of the parameters in the training of the single-layered digital neural network. (d) The training accuracies versus iterations. (e) The testing accuracies versus iterations.

Note 19. Recognition of lung diseases based on Chest X-ray images

The full NIH Chest X-ray-8 dataset is comprised of 112,120 1,024 × 1,024 high-resolution Chest X-ray (CXR) images in 8 classes. The dataset used in the experiment can be found at https://www.cc.nih.gov/drd/summers.html. The CXR images are labeled with one of the following 8 disease classes: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax, as shown in Figure S19a. 2,000 images are extracted from the full dataset to create a new dataset for the experiment. The category distribution of the new dataset for the dataset is illustrated in Figure S19b. The dataset is randomly split into the training dataset and testing dataset with a ratio of 75 % : 25%, i.e., 1,500 images for training and 500 images for testing. The category distributions of the training dataset and testing dataset are shown in Figure S19b. In the experiment, the original images with a pixel scale of 1,024 × 1,024 are encoded into the SLM to generate the optical images. After that, a 4-*f* optical system is placed between the SLM and the metasurface to minify double the area of the optical image in the optical domain, thus the optical image can effectively overlap with the metasurface chip. Followed by that, the camera is used to capture the images processed by the MOLM. The captured images are downsampled to a pixel scale of 30 × 40 using the method introduced in **Note 15**. Figure S19c lists the parameters for downsampling camera-captured images. At last, the downsampled images are used to train a single-layered digital neural network whose weight number is 1200 × 8 = 9600. Logistic regression is used to optimize the digital weights of the neural network by iterations. Considering that the pneumonia-labeled images are much less than others, we intentionally increase the training images of this category to 288 from 57 by directly copying the experimental images. This step realizes label balance which is helpful to obtain a better classification accuracy. As shown in Figures S19d − e, the training and testing accuracies are 88.7% and 85.4 %, respectively.

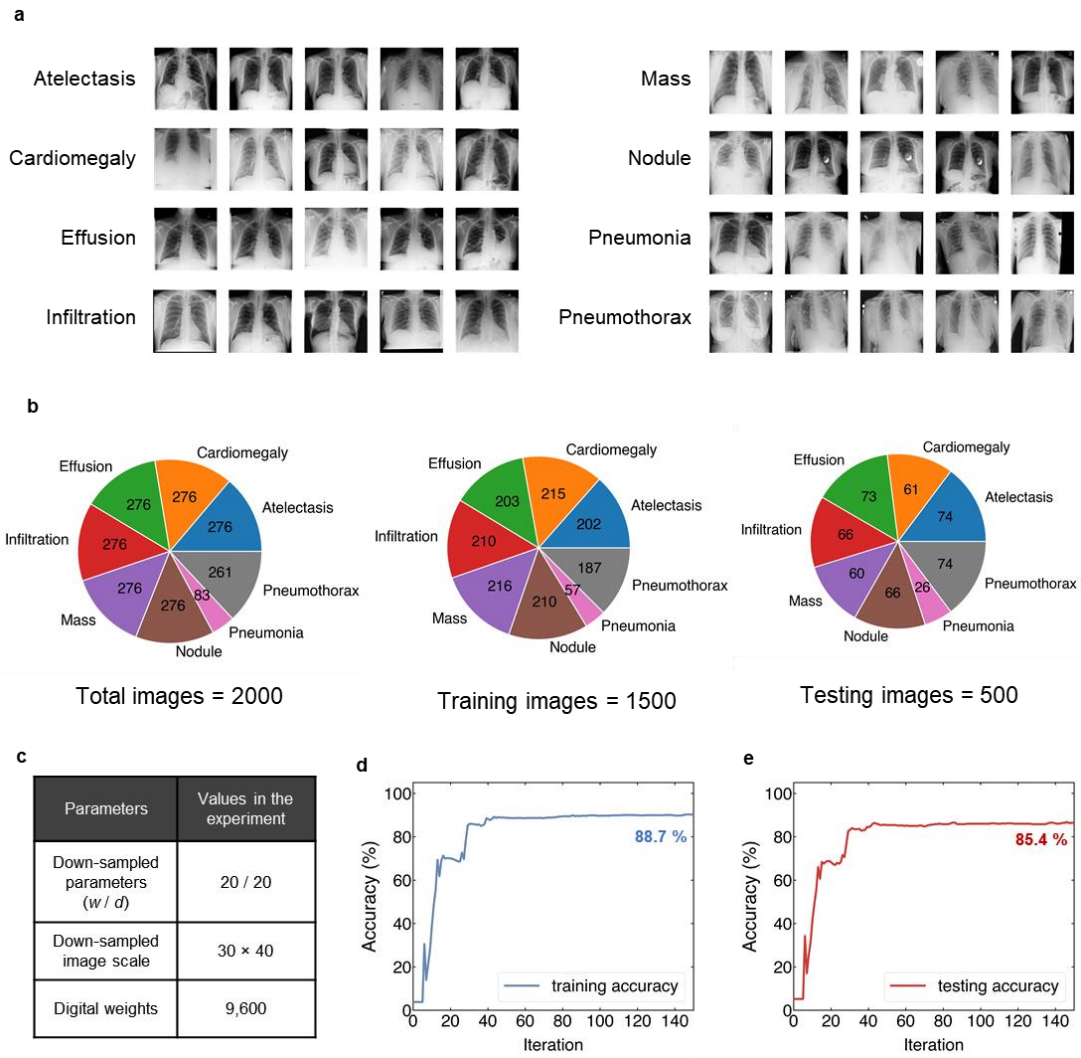


Figure S19. (a) The illustration of the NIH Chest X-ray-8 dataset. (b) The category distribution of the full dataset, training dataset, and testing dataset. The number in the pie chart indicates the image number of every category. (c) The summary table of the parameters in the training of the single-layered digital neural network. (d) The training accuracies versus iterations. (e) The testing accuracies versus iterations.

Note 20. Breast cancer detection based on whole slide image (WSI)

The full CAMELYON16 dataset consists of 400 whole-slide images (WSIs) of sentinel lymph nodes, each of which has a total pixel number of > 2.2 billion at the highest resolution. The dataset in the experiment can be found at https://camelyon16.grand-challenge.org/Home/. These WSIs are divided into a training dataset and a testing dataset, where the training dataset is comprised of 270 WSIs (160 normal slides and 110 slides containing tumor cell regions). As shown in Figure S20a, two tumor cell regions are indicated by the blue boundaries. The goal of the task is to locate the region containing the tumor cells from the billion-pixel image by our MOLM. In the experiment, one WSI image with 97,792 × 217,088 pixels is used for training our MOLM, another WSI image with 97,792 × 221,184 pixels is used for testing, as shown in Figure S20b. Considering that the WSIs cannot be directly processed by the MOLM due to the extremely large pixel number, we crop the WSI into many patch images each with a size of 1,000 × 1,000 pixels and process these patches one by one. The training WSI consists of 1,775 normal patches and 887 tumor patches. The testing WSI consists of 1,505 normal patches and 525 tumor patches. The training patches are randomly split into a new training dataset and an additional validation dataset with a split ratio of 90 % : 10 %, where the validation dataset is to independently evaluate the performance of the neural network during the training status. The category distribution of the patches of the training and validation dataset are given in Figure S20c. In the experiment, the patch images with a pixel scale of 1,000 × 1,000 are encoded into the SLM to generate the optical images. After that, a 4-*f* optical system is placed between the SLM and the metasurface to minify double the area of the optical image in the optical domain, thus the optical image can effectively overlap with the metasurface chip. Followed by that, the camera is used to capture the images processed by the MOLM. The captured images are downsampled to a pixel scale of 30 × 40 using the method introduced in **Note 15**. Figure S20d lists the parameters for downsampling camera-captured images. At last, the downsampled images are used to train a single-layered digital neural network whose weight number is 1,200 × 1 = 1,200. Logistic regression is used to optimize the digital weights of the neural network by iterations. As shown in Figures S20e–f, the training and validation accuracies are 97.2 % and 95.1 % respectively. The unlabeled patches are processed by our MOLM and then

fed into the model to produce tumor-positive probabilities. At last, these patch-level classification probabilities are mapped to the raw WSI to form a heat map with the diagnosed locations of tumor cells, where a threshold of 75 % is set to improve visualization.

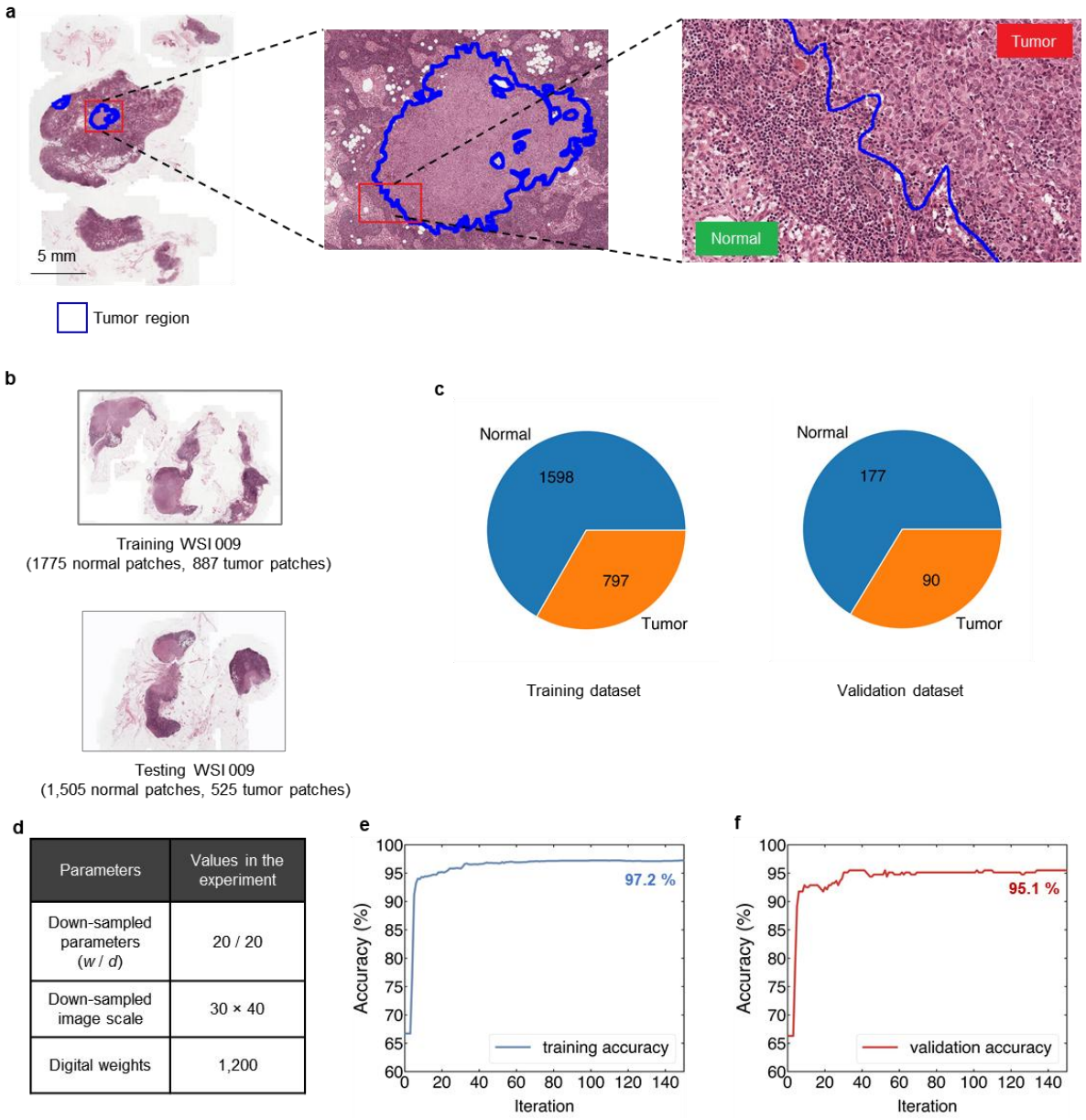


Figure S20. (a) The illustration of the CAMELYON16 dataset. (b) The illustration of the training and testing WSIs. (c) The category distribution of the training and testing patches. The number in the pie chart indicates the image number of every category. (d) The summary table of the parameters in the training of the single-

layered digital neural network. (e) The training accuracies versus iterations. (f) The validation accuracies versus iterations.

Note 21. Human action recognition based on video streams

The full KTH dataset contains 6 types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 persons in four different scenarios. Every action is repeated 4 times by one person. The action sequences are captured in the form of videos with a static camera with a 25-fps frame rate and a pixel scale of 120 × 160. The dataset in the experiment can be found at https://www.csc.kth.se/cvap/actions/. The person in the image of the original dataset is aligned to the center of the scene to create a new dataset for the experiment. Note that, except for the alignment, we do not preprocess the image using binary masking or Histogram of Oriented Gradient (HOG) as other works performed [5–6]. Figure S21a shows the action sequences of the original dataset and experimental dataset. The dataset for the experiment is extracted from the outdoor scenario $s1$ of the full KTH dataset, so the total number of action sequences is 25 × 6 × 4 = 600. The frames are extracted with an interval of 1, so the time interval between the adjoint frames is 80 ms. Every action sequence consists of 20 frames in total. The dataset is randomly split into a training dataset and a testing dataset based on persons with a ratio of 16 : 9. The training dataset consists of 384 action sequences (7680 frames) performed by 16 persons, where the frame number of every action category is 1280. The testing dataset consists of 216 action sequences (i.e., 4320 frames) performed by 9 persons, where the frame number of every action category is 720. In the experiment, the frames are upsampled to a pixel scale of 512 × 512 and then loaded to the SLM to generate optical images. The purpose of upsampling is to ensure that the SLM-generated optical image can effectively overlap with the metasurface chip. Followed by that, the camera is used to capture the images processed by the MOLM. The captured images are downsampled to a pixel scale of 34 × 45 using the method introduced in **Note 15**. Figure S21b lists the parameters for downsampling camera-captured images. At last, the downsampled images are used to train a single-layered digital neural network whose weight number is 1,530 × 6 = 9,180. Logistic regression is used to optimize the

digital weights of the neural network by iterations. As shown in Figures S21c–d, the training and testing frame-wise accuracies reach 98.5 % and 97.5 % using only 150 iterations, respectively. The frame-wise accuracy of 97.5 % is corresponding an action accuracy of 99.1 % using the most voting approach.

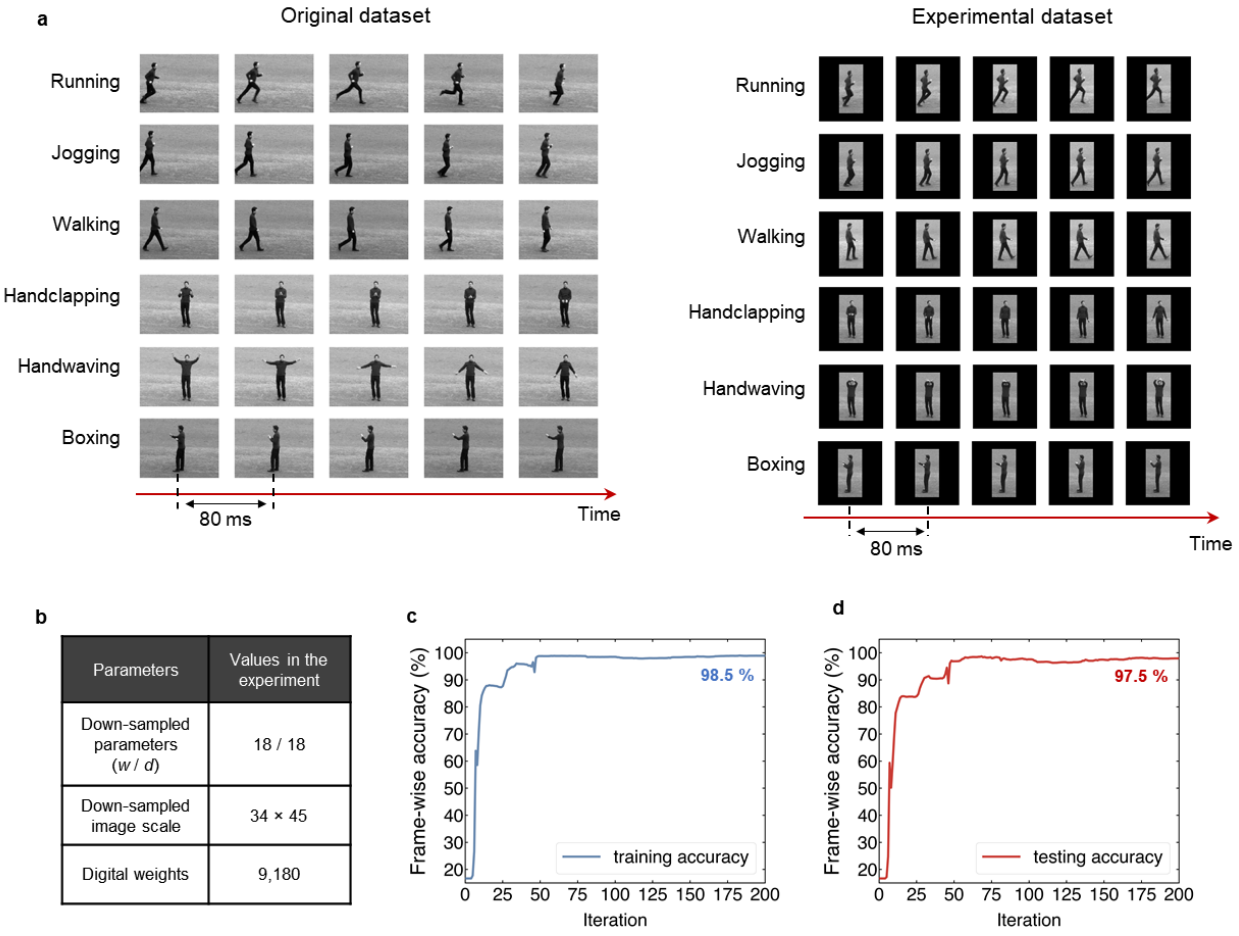


Figure S21. (a) The illustration of the original KTH dataset and the experimental dataset. (b) The summary table of the parameters in the training of the single-layered digital neural network. (c) The training frame-wise accuracies versus iterations. (d) The testing frame-wise accuracies versus iterations.

Note 22. Digital neural networks setup and training for SAM, ViT, ResNet-50 model

This section provides further details on the architecture of the digital neural network, data preprocessing, and network training for image classification tasks.

Regarding the architecture of the digital neural networks, we have chosen two transformer-based networks and one convolutional neural network. The first network is SAM [7], which combines convolutional neural networks and transformer-based architectures to process images hierarchically and at multiple scales. The second network is ViT [8], which consists of an encoder with multi-head self-attention layers, multi-layer perceptron layers, and layer normalization layers. Lastly, we utilize ResNet50 [9], a deep convolutional neural network with residual connections. While the transformer-based networks (SAM and ViT) focus on capturing global context and long-range connections, enabling them to attend to any part of the image directly and integrate information from distant pixels in a single operation, ResNet50 emphasizes leveraging local features through stacked convolutional layers. The architectural innovation of ResNet50, the residual connection, enables training of deeper networks by facilitating the propagation of the training signal, making it a powerful backbone for various computer vision tasks.

For the data preparation of benchmark datasets, we have constructed datasets consisting of various types of images, including Brain-ICH [10], Chest X-ray-8 [11], COVID-19 [12], and Camelyon16 [13]. For a fair comparison with our MOLM, we used the same method as in **Notes 16–21** to prepare the dataset. To enhance model training, we apply data augmentation techniques, including random horizontal and vertical flips with a probability of 0.5 and random rotation with 90 degrees. Regarding the input image size, the input sizes are 1024 × 1024 for SAM, 224 × 224 for ViT, and 512 × 512 for ResNet50.

As for the digital network training, we utilize the PyTorch library for implementation, and the network training is performed on the GPU of an NVIDIA RTX 3090. The networks are trained in a supervised manner, where predictions are supervised by the corresponding classification labels. We initialize the networks using weights pre-trained on ImageNet21K [14] and employ the Adam optimizer with a learning rate of 5×10^{-5} for training.

A batch size of 32 is chosen, and the maximum number of epochs is set to 200 to ensure model convergence during training.

Note 23. Time consumption of the MOLM during training and inference phases

In this section, details about the calculation of the training time and the inference time of the MOLM will be introduced. Table S1, Table S2, and Table S3 give the details about the calculation of training time and inference time of the MOLM in 4 benchmark tasks, tumor detection based on whole slide images, and human action recognition based on video streams, respectively.

The training time refers to the time consumed to train a digital neural network using the raw digital training dataset. The details of neural network training are introduced in **Note 15**. The inference time refers to the time consumed to generate the classification result using a raw digital image. The total computing time include the optical signal generation, signal propagation, detection, and computing the digital neural networks. The optical generation time is determined by the SLM response time, which is 16.7 ms per frame in the current experimental setup. Nevertheless, this optical generation time can easily be shortened to 500 μs per frame using a commercial SLM with a frame rate of 2 kHz and a pixel scale of 1024 × 1024 [15]. This value could be further improved to 1 ns per frame using a state-of-the-art SLM reported in the reference [16] or VSCEL arrays. The signal propagation time is calculated as 1.7 ns by the propagation distance (50 cm in the experiment) divided by the light speed. This value can be improved to 34 ps obtained from a propagation distance of 1 cm upon a more compact system volume. For the signal detection, the time consumption is from the response time of the camera. The response time of the camera is 16.7 ms per frame in the current experimental setup. Nevertheless, this response time can be shortened to 100 ns per frame using a commercial high-speed camera [17] or a state-of-the-art camera reported in the reference [18]. For computing the digital neural network, we consider the time spent on image downsampling, training and testing the digital neural network, and frame fusion (only in the video-based task), respectively. Image downsampling can be realized in the future using the pixel binning of the camera itself in the electrical analog domain, or using an optical 4-*f* optical system to

shrink the optical images in the optical domain, so this value can be improved to 0. The time for frame fusion refers to the time consumed for summing the current frame and the previous frame in a recurrent neural network in human action recognition based on video streams. The above time consumption in the digital domain is experimentally obtained using an NVIDIA RTX 3090-based computer. All the digital algorithms used for MOLM are based on TensorFlow 2.8 and Python 3.9. The time consumption of digital models (SAM, ResNet-50, and ViT) is experimentally obtained using an NVIDIA RTX 3090-based computer. Table S4 shows the comparison of training time and inference time of the MOLM and three types of digital models in 4 benchmark tasks. With the commercial high-speed SLM and camera, our metasurface can realize a reduction of 10^5 times in training time and 10 to 100 times in the inference time, compared with the digital models. With the state-of-the-art SLM and camera, the reduction of inference time could be further promoted to 10^3 to 10^5 times.

Parameters	Training time			Inference time per an image		
Benchmark tasks	COVID-19	Chest-X-ray-8	Brain ICH	COVID-19	Chest-X-ray-8	Brain ICH
SLM response	25.05 s / 250 ms† / 1.5 μs*			16.7 ms / 500 μs† / 1 ns*		
Free-space propagation	2.55 μs / 0.051 μs†*			1.7 ns / 0.034 ns†*		
Camera response	25.05 s / 150 μs†*			16.7 ms / 100 ns†*		
Image downsampling	1.5 s / 0†*			1 ms / 0†*		
Backend digital neural network	11 ms	350 ms	18 ms	0.41 μs	3.07 μs	0.50 μs
Total time consumption (current experiment with slow-speed SLM and camera)	51.61 s	51.95 s	51.62 s	34.4 ms	34.4 ms	34.4 ms
Total time consumption (future improvement with commercial high-speed SLM and camera)	260.15 ms	600.15 ms	268.15 ms	500.5 μs	503.2 μs	500.6 μs
Total time consumption (future improvement with the state-of-the-art SLM and camera)	11.15 ms	350.15 ms	18.15 ms	0.51 μs	3.17 μs	0.60 μs

Table S1. Details about the calculation of the training time and the inference time of the MOLM in 4 benchmark tasks. '†' and '⋆' represent the value obtained using the commercial high-speed and the state-of-the-art SLM and camera, respectively. Other parameters are experimentally obtained.

Task	Breast cancer detection based on whole slide image	
Parameters	Training time	Inference time for a WSI
SLM response	44.5 s / 1.33 s† / 2.66 μs⋆	33.9 s / 1.015 s† / 2.03 μs⋆
Free-space propagation	4.53 μs / 0.091 μs†⋆	3.45 μs / 0.069 μs†⋆
Camera response	44.5 s / 0.27 ms†⋆	33.9 s / 0.203 ms†⋆
Image downsampling	2.66 s / 0†⋆	2.03 s / 0†⋆
Backend digital neural network	125.9 ms	4.9 ms
Total time consumption (current experiment with slow-speed SLM and camera)	91.79 s	69.83 s
Total time consumption (future improvement by commercial high-speed SLM and camera)	1.46 s	1.02 s
Total time consumption (future improvement by the state-of-the-art SLM and camera)	126.2 ms	6.1 ms
Time consumption (SAM on NVIDIA GTX3090)	37.8 h	1.48 h

Table S2. Details about the calculation of the training time and the inference time of the MOLM in breast cancer detection based on whole slide image. '†' and '⋆' represent the value obtained using the commercial high-speed and the state-of-the-art SLM and camera, respectively. Other parameters are experimentally obtained.

Task	Human action recognition based on video streams	
Parameters	Training time	Inference time for a video frame
SLM response	128.3 s / 3.84 s† / 7.68 μs*	16.7 ms / 0.5 ms† / 1 ns*
Free-space propagation	13.06 ms / 0.26 ms†*	1.7 ns / 0.034 ns†*
Camera response	128.3 s / 0.77 ms†*	16.7 ms / 100 ns†*
Image downsampling	7.68 s / 0†*	1 ms / 0†*
Frame fusion	14.7 ms	1.92 μs
Backend digital neural network	4.0 s	6.3 μs
Total time consumption (current experiment with slow-speed SLM and camera)	268.3 s	34.4 ms → 29 fps
Total time consumption (future improvement with commercial high-speed SLM and camera)	7.84 s	0.508 ms → 1,968 fps
Total time consumption (future improvement with the state-of-the-art SLM and camera)	4.01 s	8.32 μs → 120,192 fps

Table S3. Details about the calculation of the training time and the inference time of MOLM in human action recognition based on video streams. '†' and '*' represent the value obtained using the commercial high-speed and the state-of-the-art SLM and camera, respectively. Other parameters are experimentally obtained.

Tasks	COVID-19		Chest-X-ray-8		Brain ICH	
Computing hardware	Training time (s)	Inference time (μs)	Training time (s)	Inference time (μs)	Training time (s)	Inference time (μs)
SAM on NVIDIA RTX 3090	53,000 s	203.54 ms	53,700 s	192.33 ms	57,200 s	205.98 ms
ResNet-50 on NVIDIA RTX 3090	32,000 s	21.83 ms	27,900 s	15.08 ms	41,800 s	23.75 ms
ViT on NVIDIA RTX 3090	35,500 s	22.75 ms	26,300 s	14.97 ms	57,000 s	20.13 ms
Meta-ONN (current experiment with slow-speed SLM and camera)	51.61 s	34.4 ms	51.95 s	34.4 ms	51.62 s	34.4 ms
Meta-ONN (future improvement with commercial high-speed SLM and camera)	260.15 ms	502.1 μs	600.15 ms	502.1 μs	268.15 ms	502.1 μs
Meta-ONN (future improvement with the state-of-the-art SLM and camera)	11.15 ms	0.51 μs	350.15 ms	3.17 μs	18.15 ms	0.60 μs

Table S4. Comparison of training time and inference time per image of the MOLM and digital models in 4 benchmark tasks.

Note 24. Energy consumption of the MOLM during training and inference phases

In this section, the details about the calculation of the total energy consumption during the training and inference phases will be introduced. The total energy consumption is the sum of the energy consumption of every device in the experimental setup. The energy consumption of the device is calculated by the power consumption of the device multiplied by running time. The power consumption of the MOLM is from the laser source, SLM, camera, and the digital computer, respectively, so the total energy consumption of our MOLM is calculated using the following formula:

$$E_{total} = P_L(t_{SLM} + t_p + t_c) + P_{SLM}t_{SLM} + P_c t_c + P_{digital}t_{digital}$$

where the description and value of every expression are given in Table S5. Note that the power consumption of the digital computer based on a GPU (NVIDIA RTX 3090) contains the power consumption of all the electronic devices (the motherboard, memory, cooling system, etc.). For the commercial high-speed SLM and camera, the power consumption used for the calculation of energy consumption is the rated power consumption

obtained from the specification of the product. For the state-of-the-art SLM, the power consumption of the high-speed SLM is calculated by the power consumption of the single device multiplied by the number of the total pixels used in the experiment. According to the reference [16], the power consumption per pixel can be 1 mW, so the total power consumption of the SLM with a pixel scale of 512 × 512 is around 262 W. For the state-of-the-art camera, the power consumption of the high-speed camera is 10 W obtained from the reference [18]. The running time of every device during the training and inference phases is given in Table S1.

Table S6 gives the comparison of the energy consumption during the training and inference phases of the MOLM and digital models in 4 benchmark tasks. For the results obtained by the experiment with commercial equipment, our MOLM achieves a reduction in total energy consumption by 10^4 times during the training phase and 2 to 100 times during the inference phase, compared with digital models. With the commercial high-speed SLM and camera, the reduction of energy consumption can be promoted to 10^4 to 10^6 times during the training phase and 10^2 to 10^3 during the inference phase. With the state-of-the-art SLM and camera, the reduction of energy consumption in the inference time could be further promoted to 10^3 to 10^5 times.

Expression	Description	Value
P_L	The power consumption of the laser diode	5 W
P_{SLM}	The power consumption of the SLM	12 W / 12 W† / 262 W*
P_c	The power consumption of the camera	2.5 W / 200 W† / 10 W*
$P_{digital}$	The power consumption of the GPU-based digital computer (including the motherboard, memory, cooling system, etc.)	750 W
t_{SLM}	The running time of the SLM	The value of row "SLM response" in Table S1
t_p	The time of the free-space propagation of the light	The value of row "Free-space propagation " in Table S1
t_c	The running time of the camera	The value of row "Camera response" in Table S1
$t_{digital}$	The running time of the digital computer	For meta-ONN, the value is the sum of the value of row "Image downsampling" and "Backend digital neural network" in Table S1. For digital models, the value is given in Table S4.

Table S5. The summary of the description and value of every expression of the formulation for calculating the total energy consumption. '†' and '⋆' represent the value obtained using the commercial high-speed and the state-of-the-art SLM and camera, respectively.

Tasks	COVID-19		Chest-X-ray-8		Brain ICH	
Total energy consumption (E_{total})	Training (J)	Inference (mJ)	Training (J)	Inference (mJ)	Training (J)	Inference (mJ)
SAM	3.98×10^7	1.53×10^5	4.03×10^7	1.44×10^5	4.29×10^7	1.54×10^5
ResNet-50	2.40×10^7	1.64×10^4	2.09×10^7	1.13×10^4	3.14×10^7	1.78×10^4
ViT	2.66×10^7	1.71×10^4	1.97×10^7	1.12×10^4	4.28×10^7	1.51×10^4
Meta-ONN (current experiment with slow-speed SLM and camera)	1.75×10^3	1,159	2.0×10^3	1,161	1.75×10^3	1,160
Meta-ONN (future improvement with commercial high-speed SLM and camera)	12.5	8.83	267	108.2	17.8	8.90
Meta-ONN (future improvement with the state-of-the-art SLM and camera)	8.25	0.31	263	2.30	13.5	0.38

Table S6. Comparison of the energy consumption during training and inference phases of the MOLM and digital models in 4 benchmark tasks.

Note 25. TSNE analysis for experimentally captured images

T-distributed Stochastic Neighbor Embedding (TSNE) is a tool to visualize high-dimensional data. In the manuscript, we use the code package [19] to visualize the images after the MOLM (i.e., after the camera and before the digital signal-layer NN). The pixel scale of the input image to TSNE is 300 × 400, obtained by down-sampling the original experimental images after normalizing the intensity to 0 to 1. The number of images sent to TSNE for evaluation is 500, which are randomly chosen from all the experimental images.

Note 26. Explanation of the enhancement of information entropy and high-frequency components

In this section, we conduct simulations to elaborate on why increasing the number of meta-atoms enhances the information entropy and spatial frequency. We first introduce the simulation settings. Then, we interpret the results to explain the enhancement mechanism.

Simulation settings: In our simulations, the physical dimensions of both the metasurface chip and the detector array are fixed at 3.2 mm × 3.2 mm. The effective number of independently modulated meta-atoms is varied

from 160,000 to 41 million. To ensure that these settings are physically valid, we model the metasurface on a fixed 6400 × 6400 lattice of meta-atoms with a 500 nm period. The effective node number is then adjusted by grouping neighboring meta-atoms and let them share the same modulation coefficient; i.e., equivalently, by changing the spatial density of the modulation coefficients (see Figure S22a).

The simulation starts with an MNIST optical image of a "0," which is element-wise modulated by both the phase and amplitude coefficients of the metasurface chip. The modulated optical image then propagates a distance of 3.2 mm to the detector array plane. The output image is the modulus square of the optical field on the detector array surface. The simulation wavelength is 532 nm, and the simulations are performed using the angular spectrum diffraction method [20].

Simulation results: We increase the effective number of meta-atoms from 160,000 to 41 million, which reflects the increased spatial density of the phase coefficient pattern, as shown in Figure S22a. As the effective number of meta-atoms increases, the speckle distribution of the output optical field becomes denser in space, as shown in Figure S22b. By combining the simulated results with the theoretical derivation, the mechanism behind the enhancement of high-frequency components and information entropy is explained as follows:

1. Enhancement of high-frequency components

Before presenting the simulation results, we mathematically analyze how the phase coefficient pattern enhances the high-frequency components of the output image. When an input image $\boldsymbol{X}$ is element-wise modulated by a phase coefficient pattern $\boldsymbol{W}$, the output image can be expressed as $\boldsymbol{X}\odot\boldsymbol{W}$, where $\odot$ denotes element-wise multiplication. The spatial frequency spectrum of the output image can be further expressed as $\mathcal{F}\{\boldsymbol{X}\odot\boldsymbol{W}\} = \mathcal{F}\{\boldsymbol{X}\}\otimes\mathcal{F}\{\boldsymbol{W}\}$, where $\mathcal{F}\{\cdot\}$ and $\otimes$ denote the 2-D Fourier transform and 2-D convolution, respectively [21]. The frequency spectrum of the input image, $\mathcal{F}\{\boldsymbol{X}\}$, is mostly concentrated in the low-frequency region, while the frequency spectrum of the phase coefficient pattern, $\mathcal{F}\{\boldsymbol{W}\}$, is mostly concentrated in the high-frequency region. As a result, convolving $\mathcal{F}\{\boldsymbol{X}\}$ with $\mathcal{F}\{\boldsymbol{W}\}$ smears the low-frequency energy of $\mathcal{F}\{\boldsymbol{X}\}$ to the high-frequency region, effectively redistributing energy from low to high frequencies. As the high-

frequency region of $\mathcal{F}\{\boldsymbol{W}\}$ expands (i.e., as the spatial density of the phase coefficient pattern increases), more low-frequency components are spread into higher-frequency regions.

Our simulation results further support the above conclusion. As the effective number of meta-atoms increases (i.e., as the spatial density increases), more low-frequency components are transferred to higher-frequency regions, as shown in Figure S22c. As a result, the high-frequency features are overall enhanced at 41 million meta-atoms. The theoretical connection between enhancing high-frequency components and machine learning capability is explained in Ref. [22].

2. Elevation of the information entropy

For an information carrier whose element values follow a standard Gaussian distribution, the information entropy is given by $\frac{1}{2} ln\ (2\pi e\sigma^2)$, where σ^2 is the variance of the distribution [23]. This indicates that a broader distribution yields higher information entropy.

In our system, the Gaussian-random phase coefficient enables the pixel values of the output speckle pattern to also approximately follow a Gaussian distribution (see Figure S22d). As the effective number of meta-atoms increases, the distribution of pixel values broadens, leading to a larger variance. Consequently, the information entropy of the output speckle pattern increases, from 7.2 bits at 160,000 meta-atoms to 11.1 bits at 41 million meta-atoms, as shown in Figure S22e. The information entropy is calculated using the standard formulation $-\sum_{k=1}^{M} \widetilde{U}_k log_2(\widetilde{U}_k)$, where $\widetilde{U}_k = U_k / \sum_{k=1}^{M} U_k$, U_k denotes the k th pixel value, and M is the total number of pixels.

With the enhancement of high-frequency components and information entropy, the MNIST classification accuracy improves from 91.0 % at 160,000 effective meta-atoms to 94.7 % at 41 million effective meta-atoms, as shown in Figures S22f and S22g. To study the effect of the meta-atom number independently, the above MNIST accuracies are obtained using a metasurface without an optical lens. Therefore, the MNIST accuracy exhibits a slight difference from the experimental result (99.2 % at 41 million meta-atoms with an optical lens).

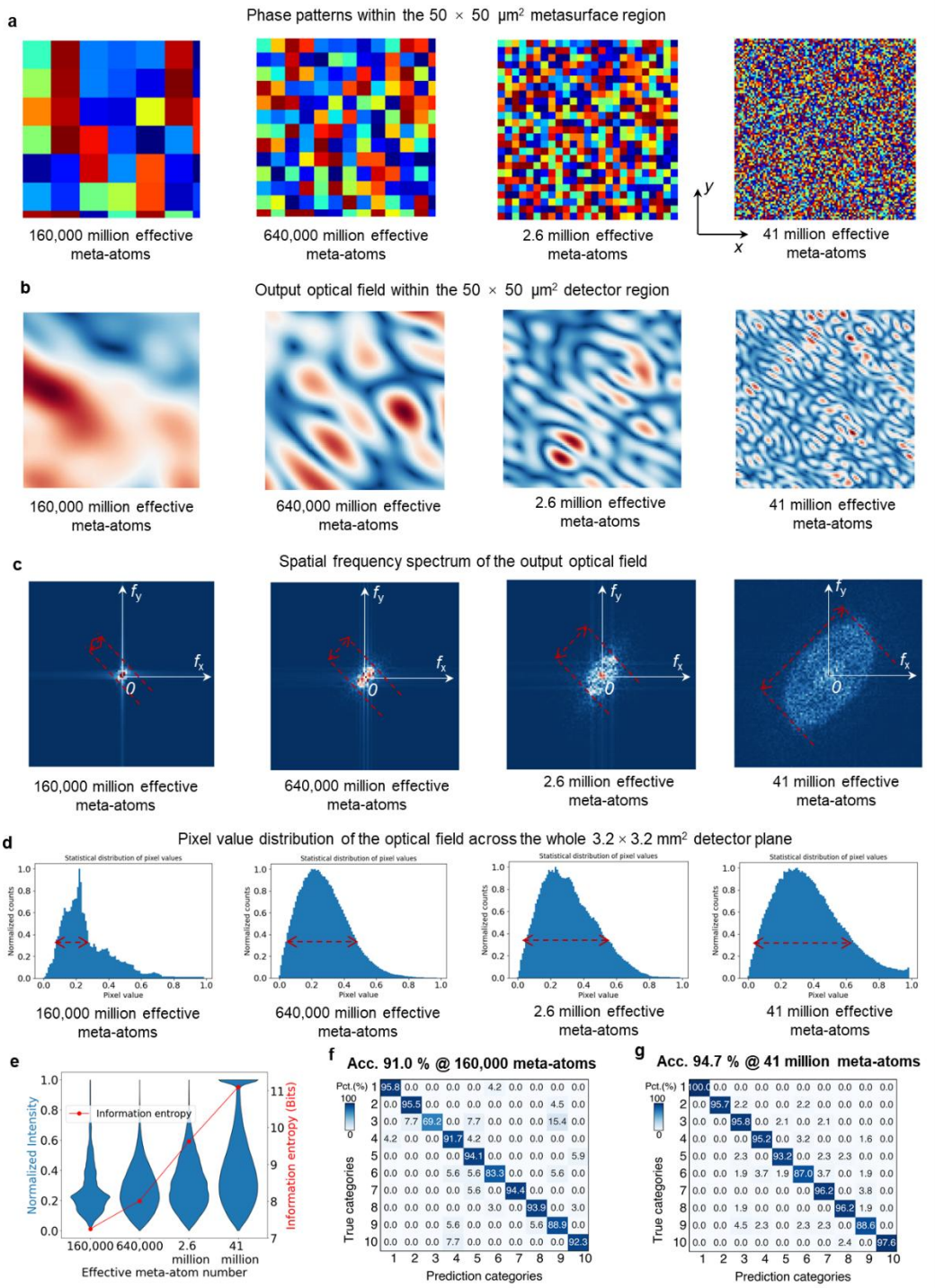


Figure S22. (a) Phase modulation patterns for different numbers of effective meta-atoms. (b) Output optical field for different numbers of effective meta-atoms. (c) Spatial frequency spectrum of the output optical field for different numbers of effective meta-atoms. (d) Statistical distribution of pixel values in the output optical field for different numbers of effective meta-atoms. (e) Information entropy versus the number of effective meta-atoms. (f) The confusion matrix of the MNIST classification result for 160,000 meta-atoms. (g) The confusion matrix of the MNIST classification result for 41 million meta-atoms.

Note 27. Performance comparison tables

References	Implementation of optical frontends	Randomness of optical parameters	MNIST accuracy	CIFAR-10 accuracy	Digital Params
[24]	On-chip photonic devices	Controllable but not explicitly designed	91.6 %	N/A	2,560
[25]			94.5 %		5,000
[26]	Time-delayed laser loop	Hard to control	97.4 %	N/A	10,000
[27]			96.3 %		
[28]	Multi-mode optical fiber	Hard to control	96.8 %	N/A	>1 million
[29]			94.3 %		20,250
[30]			96.0 %		70,000
[31]			89.3 %		1,500
[32]			94.8 %		1,440
[33]	Random scattering media	Hard to control	97.9 %	N/A	>600,000
[34]			98.0 %	N/A	100,000
[35]			96.0 %	49.2 %	35,000
[36]			N/A	79.2 %	>60 million
[37]	Spatial light modulator	Controllable but not explicitly designed	92.1 %	N/A	40,960
[38]			97.0 %		160,000
Ours	Optical metasurface	Explicitly designed	99.2 %	91.6 %[1]	3,000

Table S7. The performance comparison of our MOLM with other ONN systems with untrained optical frontends. [1]The accuracy is obtained using 7,440 trained digital parameters.

References	Methods	Dataset	Processed data size[2]	# of optical parameters	Task type	Task performance
[39]	2D integrated photonics	FNC[1]	1×1	8	Classification	N/A
[40]		Vowel	1×4	16		76.7%
[41]		EMNIST	3×4	9		89.9%
[42]		MNIST	28×28	16		95.3%
[43]		CIFAR-10	32×32	7		79.8%[3]
[44]	3D free-space photonics	MNIST	28×28	2×10^5	Classification	91.75%
[45]		CIFAR-10	32×32	3.9×10^4		73.12%
[46]		KTH	120×160	4.9×10^5		96.3%
[47]		KTH	120×160	1.6×10^3		91.3%
[48]	Multimode optical fiber	COVID-19	299×299	N/A	Classification	83.2%
[49]		COVID-19	299×299	N/A		77%
[50]	Optical metasurface	MNIST	28×28	7.84×10^4	Classification	93.75%[4]
[51]		MNIST	28×28	1.25×10^4		98.05%
[52]		MNIST	28×28	6.27×10^5		93.1%
[53]		MNIST	28×28	3.39×10^5		98.6%
Ours	Optical metasurface	MNIST	28×28	4.1×10^7	Classification	99.2%
		KTH	120×160			99.1%
		COVID-19	299×299			97.0%
		Brain ICH	512×512			97.8%
		ChestX-ray8	$1{,}024 \times 1{,}024$			85.4%
		Brain ICH	512×512		Localization	IOU=0.61
		CAMELYON	2 billion pixels[5]		Segmentation	IOU=0.60

Table S8. The performance comparison of our MOLM with the state-of-the-art ONN systems. [1]FNC represents the optical communication data for fiber nonlinearity compensation. [2]The processed data size refers to the

original data size of the dataset, instead of the data size that can be loaded into the ONN at one time. [3]The accuracy is experimentally obtained based on the two-categorized CIFAR-10 dataset, and the image is preprocessed by an electrical convolutional neural network before being fed into the ONN. [4]The accuracy is experimentally obtained based on the four-categorized MNIST dataset. [5]The whole slide image (WSI) from the CAMELYON-16 dataset has ~ 2 billion pixels, which are cropped into many sub-images with 1,000 × 1,000 pixels that our system can process in one shot. # represents number.

References

[1] Zhu, Long, and Jian Wang. "Arbitrary manipulation of spatial amplitude and phase using phase-only spatial light modulators." Scientific reports 4.1 (2014): 7441.

[2] Nie, Yujie, et al. "Transmission-matrix quantitative phase profilometry for accurate and fast thickness mapping of 2D materials." ACS Photonics 10.4 (2023): 1084-1092.

[3] Teğin, Uğur, et al. "Scalable optical learning operator." Nature Computational Science 1.8 (2021): 542-549.

[4] Oguz, I., Hsieh, J. L., Dinc, N. U., Teğin, U., Yildirim, M., Gigli, C., Moser, C. & Psaltis, D. “Programming Nonlinear Propagation for Efficient Optical Learning Machines.” arXiv preprint arXiv:2208.04951(2022)

[5] Zhou, Tiankuang, et al. "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit." Nature Photonics 15.5 (2021): 367-373.

[6] Antonik, Piotr, et al. "Human action recognition with a large-scale brain-inspired photonic computer." Nature Machine Intelligence 1.11 (2019): 530-537.

[7] Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[8] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." International Conference on Learning Representations, 2021.

[9] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[10] Flanders, Adam E., Luciano M. Prevedello, George Shih, Safwan S. Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T. Mongan et al. "Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge." Radiology: Artificial Intelligence 2, no. 3. 2020.

[11] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097-2106. 2017.

[12] Rahman, Tawsifur, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam et al. "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images." Computers in biology and medicine 132, 2021.

[13] Bejnordi, Babak Ehteshami, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." Jama 318, no. 22. 2017.

[14] Ridnik, Tal, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. "Imagenet-21k pretraining for the masses." Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021.

[15] https://www.sintec.sg/optics/380.html

[16] Smolyaninov, Alexei, et al. "Programmable plasmonic phase modulation of free-space wavefronts at gigahertz rates." Nature Photonics 13.6 (2019): 431-435.

[17] https://www.shimadzu.com/an/products/materials-testing/high-speed-video-camera/hyper-vision-hpv-x2/spec.html

[18] Kuroda, Rihito, et al. "A 20Mfps global shutter CMOS image sensor with improved light sensitivity and power consumption performances." ITE Transactions on Media Technology and Applications 4.2 (2016): 149-154.

[19] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

[20] Matsushima, Kyoji, Hagen Schimmel, and Frank Wyrowski. "Fast calculation method for optical diffraction on tilted planes by use of the angular spectrum of plane waves." *Journal of the Optical Society of America A* 20.9 (**2003**): 1755-1762.

[21] Hansen, Eric W. *Fourier transforms: principles and applications*. John Wiley & Sons, **2014**.

[22] Tancik, Matthew, et al. "Fourier features let networks learn high frequency functions in low dimensional domains." *Advances in neural information processing systems* 33 (**2020**): 7537 7547.

[23] Thomas, M. T. C. A. J., and A. Thomas Joy. *Elements of information theory*. Wiley-Interscience, **2006**.

[24] Nakajima, M., Tanaka, K., Hashimoto, T.: Scalable reservoir computing on coherent linear photonic processor. *Communications Physics* 4(1), 20 (**2021**)

[25] Lugnan, A., Foradori, A., Biasi, S., Bienstman, P., Pavesi, L.: On-chip photonic neural network for multi-wavelength time-dependent signal processing. *In: 2024 IEEE Photonics Conference (IPC)*, pp. 1–2 (**2024**). IEEE

[26] Huang, Y., Zhou, P., Yang, Y., Chen, T., Li, N.: Time-delayed reservoir computing based on a two-element phased laser array for image identification. *IEEE Photonics Journal* 13(5), 1–9 (**2021**)

[27] Zhou, C., Mu, P., Huang, Y., Yang, Y., Zhou, P., Lau, K., Li, N.: Deep photonic reservoir computing based on a distributed feedback laser array. *APL Photonics* 10(2) (**2025**)

[28] Oguz, I., Hsieh, J.-L., Dinc, N.U., Te˘gin, U., Yildirim, M., Gigli, C., Moser, C., Psaltis, D.: Programming nonlinear propagation for efficient optical learning machines. *Advanced Photonics* 6(1), 016002–016002 (**2024**)

[29] Ancora, D., Negri, M., Gianfrate, A., Trypogeorgos, D., Dominici, L., Sanvitto,

D., Ricci-Tersenghi, F., Leuzzi, L.: Low-power multimode-fiber projector outperforms shallow-neural-network classifiers. *Physical Review Applied* 21(6), 064027 (**2024**)

[30] Borhani, N., Kakkava, E., Moser, C., Psaltis, D.: Learning to see through multimode fibers. *Optica* 5(8), 960–966 (**2018**)

[31] Saeed, S., M¨uft¨uo˘glu, M., Cheeran, G.R., Bocklitz, T., Fischer, B., Chemnitz, M.: Nonlinear inference capacity of fiber-optical extreme learning machines. *Nanophotonics* (0) (**2025**)

[32] Ermolaev, A.V., Hary, M., Leybov, L., Ryczkowski, P., Skalli, A., Brunner, D., Genty, G., Dudley, J.M.: Limits of nonlinear and dispersive fiber propagation for photonic extreme learning. a*rXiv preprint arXiv:2503.03649* (**2025**)

[33] Chen, H., Gao, Y., Liu, X., Zhou, Z.: Imaging through scattering media using speckle pattern classification based support vector regression. *Optics express* 26(20), 26663–26678 (**2018**)

[34] Saade, A., Caltagirone, F., Carron, I., Daudet, L., Dr´emeau, A., Gigan, S., Krzakala, F.: Random projections through multiple optical scattering: Approximating kernels at the speed of light. *In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6215–6219 (**2016**). IEEE

[35] Wang, H., Hu, J., Morandi, A., Nardi, A., Xia, F., Li, X., Savo, R., Liu, Q., Grange, R., Gigan, S.: Large-scale photonic computing with nonlinear disordered media. *Nature Computational Science* 4(6), 429–439 (**2024**)

[36] Ohana, R., Wacker, J., Dong, J., Marmin, S., Krzakala, F., Filippone, M., Daudet, L.: Kernel computations from large-scale random features obtained by optical processing units. *In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9294–9298 (**2020**). IEEE

[37] Pierangeli, D., Marcucci, G., Conti, C.: Photonic extreme learning machine by free-space optical propagation. *Photonics Research* 9(8), 1446–1454 (**2021**)

[38] Antonik, P., Marsal, N., Rontani, D.: Large-scale spatiotemporal photonic reservoir computer for image classification. *IEEE Journal of Selected Topics in Quantum Electronics* 26(1), 1–12 (**2019**)

[39] Huang, C., Fujisawa, S., Lima, T.F., Tait, A.N., Blow, E.C., Tian, Y., Bilodeau, S., Jha, A., Yaman, F., Peng, H.-T., et al.: A silicon photonic–electronic neural network for fibre nonlinearity compensation. *Nature Electronics* 4(11), 837–844 (**2021**)

[40] Shen, Y., Harris, N.C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., et al.: Deep learning with coherent nanophotonic circuits. *Nature photonics* 11(7), 441–446 (**2017**)

[41] Ashtiani, F., Geers, A.J., Aflatouni, F.: An on-chip photonic deep neural network for image classification. *Nature* 606(7914), 501–506 (**2022**)

[42] Feldmann, J., Youngblood, N., Karpov, M., Gehring, H., Li, X., Stappers, M., Le Gallo, M., Fu, X., Lukashchuk, A., Raja, A.S., et al.: Parallel convolutional processing using an integrated photonic tensor core. *Nature* 589(7840), 52–58 (**2021**)

[43] Moralis-Pegios, M., Mourgias-Alexandris, G., Tsakyridis, A., Giamougiannis, G., Totovic, A., Dabos, G., Passalis, N., Kirtas, M., Rutirawut, T., Gardes, F., et al.: Neuromorphic silicon photonics and hardware-aware deep learning for high-speed inference. *Journal of Lightwave Technology* 40(10), 3243–3254 (**2022**)

[44] Lin, X., Rivenson, Y., Yardimci, N.T., Veli, M., Luo, Y., Jarrahi, M., Ozcan, A.: All-optical machine learning using diffractive deep neural networks. *Science* 361(6406), 1004–1008 (**2018**)

[45] Wei, K., Li, X., Froech, J., Chakravarthula, P., Whitehead, J., Tseng, E., Majumdar, A., Heide, F.: Spatially varying nanophotonic neural networks. *Science Advances* 10(45), 0391 (**2024**)

[46] Zhou, T., Lin, X., Wu, J., Chen, Y., Xie, H., Li, Y., Fan, J., Wu, H., Fang, L., Dai, Q.: Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics* 15(5), 367–373 (**2021**)

[47] Antonik, P., Marsal, N., Brunner, D., Rontani, D.: Human action recognition with a large-scale brain-inspired photonic computer. *Nature Machine Intelligence* 1(11), 530–537 (**2019**)

[48] Te˘gin, U., Yıldırım, M., O˘guz, I., Moser, C., Psaltis, D.: Scalable optical learning ˙ operator. *Nature Computational Science* 1(8), 542–549 (**2021**)

[49] Oguz, I., Hsieh, J.-L., Dinc, N.U., Te˘gin, U., Yildirim, M., Gigli, C., Moser, C., Psaltis, D.: Programming nonlinear propagation for efficient optical learning machines. *Advanced Photonics* 6(1), 016002–016002 (**2024**)

[50] Luo, X., Hu, Y., Ou, X., Li, X., Lai, J., Liu, N., Cheng, X., Pan, A., Duan, H.: Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible. *Light: Science & Applications* 11(1), 158 (**2022**)

[51] Qu, G., Cai, G., Sha, X., Chen, Q., Cheng, J., Zhang, Y., Han, J., Song, Q., Xiao, S.: All-dielectric metasurface empowered optical-electronic hybrid neural networks. *Laser & Photonics Reviews* 16(10), 2100732 (**2022**)

[52] Zheng, H., Liu, Q., Zhou, Y., Kravchenko, I.I., Huo, Y., Valentine, J.: Meta-optic accelerators for object classifiers. *Science Advances* 8(30), 6410 (**2022**)

[53] Zheng, H., Liu, Q., Kravchenko, I.I., Zhang, X., Huo, Y., Valentine, J.G.: Multichannel meta-imagers for accelerating machine vision. *Nature Nanotechnology*, 1–8 (**2024**)