Supplementary Information for

Multi-task large-scale integrated optical vision processor using ultra-fast parallel nanofabrication

Wenqi Ouyang[2,4], Wen Lyu[1,3], Jianming Xiong[1], Jiayong Peng[1], Mingcheng Luo[1], Kaifei Tang[1], Shih-Chi Chen[2,4]*, and Chaoran Huang[1]*

Corresponding author: scchen@mae.cuhk.edu.hk; crhuang@ee.cuhk.edu.hk

These authors contributed equally: Wenqi Ouyang, Wen Lyu

The PDF file includes:

Supplementary Text
Figures S1 to S8
Tables S1 to S4

Other Supplementary Materials for this manuscript include the following:

Movie S1

3D parallel nanofabrication system and fabrication strategy

We developed a system based on binary digital micromirror device (DMD) holography for randomized multi-focus two-photon lithography (TPL) nanofabrication, as shown in Supplementary Fig. S1. This platform is built around a low-repetition-rate femtosecond laser amplifier (Spitfire Pro, Spectra-Physics; central wavelength: 800 nm; repetition rate: 1 kHz; pulse width: 100 fs; average power: 4 W), with the input beam intensity adjusted via a half-wave plate and a polarizing beam splitter. The output is first diffracted by a 600-lines-per-millimeter grating and then relayed through a pair of 4f telescopes (L1 and L2) for dispersion pre-compensation before reaching the DMD (SuperSpeed V-6501, 1920 × 1080 pixels, pixel size: 7.56 μm, ViALUX). The DMD subsequently projects a binary Lee hologram whose phase pattern encodes a custom multi-focus array. A Fourier-transform lens (L3) produces the corresponding focal spot array, from which a spatial filter selects the desired diffraction order. The filtered pattern is then demagnified by a 4f relay system consisting of a lens (L4) and a high-numerical-aperture (NA) oil-immersion objective (CFI Plan Fluor 100X Oil), which focuses the femtosecond pulses into the photoresist. This is implemented in a dip-in configuration, where a droplet of photoresist on a glass substrate directly contacts the bottom of the objective lens for immersion exposure. After development, the cured thin-layer DOE structures remain adhered to the glass substrate, while the unexposed photoresist is removed. In each exposure cycle, the objective simultaneously delivers 25 focal spots into the photoresist. A yellow illumination beam is introduced from beneath the sample stage via a mirror, passing through the glass substrate to provide back illumination. The imaging shares the same oil-immersion objective used for fabrication. A longpass dichroic mirror (DM) directs the reflected yellow light through a tube lens (L5) onto a CCD camera (BFLY-U3-26S6C-C Blackfly, FLIR) for real-time monitoring. A six-axis nano-positioning stage (H-811.I2, Physik Instrumente) actively compensates for tip and tilt misalignments, which is especially critical given the ultrathin nature of the printed diffractive optical elements, and enables seamless, large-area stitching with high spatial precision.

In each exposure frame, the desired 3D intensity distribution is encoded as a complex wavefront $A(x,y)\cdot exp[i\varphi(x,y)]$, where $A(x,y)$ defines the target amplitude profile and $\varphi(x,y)$ is the corresponding phase. To generate the associated binary Lee hologram $h(i,j)$ displayed on the DMD, we implement a weighted Gerchberg-Saxton algorithm, which iteratively propagates

the optical field between the DMD and the sample planes via Fourier transforms. In each iteration, the resulting continuous phase pattern is binarized into a two-level (on/off) hologram consistent with the DMD's digital architecture. The feedback loop accounts for the quantization error introduced by binarization, enhancing focal uniformity and minimizing zero-order diffraction. Eq. S1 formulates the composite hologram as a product of the designed wavefront and an additional quadratic phase term for axial focusing. To steer the lateral position of each focal spot, a tilted grating phase $R(x,y) = (cos\,\theta{\cdot}x + sin\,\theta{\cdot}y)$ is embedded into the hologram, where θ controls the displacement direction and the grating period T (from Eq. S1) together with $R(x,y)$ defines the tilt phase for lateral beam steering. For axial positioning, we superimpose a spherical phase $\varphi(x,y) = \pi(x^2 + y^2)/(\lambda f)$, as given in Eq. S2, where λ is the laser wavelength and f is the focal length. A composite field comprising k focal spots is synthesized by summing k individually weighted holograms, each with amplitude coefficient B_k and phase offset φ_k, as expressed in Eq. S3. Here, $h(i,j)$ represents the binary state of the micromirrors on the DMD, while B_k, $R_k(x,y)$, T_k, and φ_k denote the relative amplitude factor, tilted grating phase, grating period, and phase term for the k-th focal point, respectively. φ_w, k encodes additional wavefront shaping information to control the size and shape of each focal spot. The amplitude B_k directly modulates the energy delivered to each focal point, allowing free control over voxel intensity within each single-shot exposure. Notably, this weighted Gerchberg–Saxton (WGS) algorithm achieves over 99.9% uniformity among the 25 foci after only a few iterations, ensuring consistent exposure conditions critical for high-fidelity nanofabrication.

$$h(i,j) = \begin{cases} 1, & -\frac{sin^{-1}A(x,y)}{2\pi} \le \frac{R(x,y)}{T} + \frac{\varphi(x,y)}{2\pi} + k \le \frac{sin^{-1}A(x,y)}{2\pi} \\ 0, & otherwise \end{cases} \tag{S1}$$

$$\varphi(x,y) = \frac{\pi(x^2 + y^2)}{\lambda f} \tag{S2}$$

$$h(i,j) = \begin{cases} 1, & -A(x,y) \le \sum_{k=1}^{n} B_k\, sin\left(2\pi\frac{R_k(x,y)}{T_k} + \varphi_k(x,y) + \varphi_{w,k}(x,y)\right) \le A(x,y) \\ 0, & otherwise \end{cases} \tag{S3}$$

Rather than printing voxels in a uniformly spaced and regularly arranged grid, we employ a random-access scanning strategy that generates a 3D randomly arranged multi-focus coordinate pattern for every laser pulse. Within a 50 μm × 50 μm region, 25 voxel positions are sampled at random and converted into binary holograms by the weighted Gerchberg–Saxton engine. This spatiotemporal randomization guarantees a minimum delay between exposures of neighboring voxels, effectively suppressing near-field diffusion effects, and pixel-level cross-talk. Each voxel is realized by four pulses at 1 kHz, spaced 200 nm apart along the optical axis to form a circular feature, while the six-axis stage carries out long-travel translation to cover the full 1 mm² device area. In addition, pixel heights are designed between 0 and 500 nm in 100 nm increments, with heights sampled randomly across the pattern. This axial positioning is achieved by modulating the spherical wavefront of each focus via the DMD binary hologram (mentioned above in the context of wavefront control), thereby precisely controlling the focus z position during scanning to realize accurate pixel height designs.

To evaluate the advantage of this approach, we compared it against conventional sequential scanning. In the sequential mode, printed features suffer from cumulative diffusion that blur adjacent voxels, with the blurring effect becoming more pronounced as the scan progresses. By contrast, the randomized protocol delivers pixel-level clarity, with seamlessly merged exposure regions produced by each of the 25 foci without visible boundaries or stitching artifacts. Supplementary Movie S1 captures the entire workflow, including the generation of a randomly distributed set of 500 nm pixel coordinates, the *in situ* nanofabrication under the objective, and the final SEM results of both scanning schemes that highlight the superior fidelity afforded by spatiotemporal randomization.

Randomized sub-block scanning strategy for spatially separated multi-focus exposure

To ensure both spatial separation among simultaneously exposed voxels and temporal separation between neighboring pixel-exposure events, we employ a randomized sub-block scanning strategy. First, the entire 100 × 100 pixel area (50 × 50 μm^2 at a 500 nm pixel pitch) is divided into a 5 × 5 grid of 20 × 20 pixel regions, each assigned to one of the 25 parallel laser foci. Within each region, the order of pixel exposures is randomized as described earlier to maintain pixel-level fidelity. To ensure a fully filled, gapless pixel arrangement under high-resolution fabrication, each pixel is composed of 4 vertically spaced voxels with 200 nm gaps

(denoted in Supplementary Fig. S2a), and each pixel is effectively formed by four voxels (or nine voxels for 800 nm and 1200 nm pixels).

To further enforce a minimum inter-focus distance and avoid near-field focus interference or local energy accumulation among adjacent focal spots, each 20 × 20 region is subdivided into four 10 × 10 sub-blocks (labeled A–D in Supplementary Fig. S2a). Within each sub-block, pixel coordinates are randomly permuted using a MATLAB-based pseudo-random shuffling algorithm to ensure statistically independent exposure sequences among adjacent voxels. During fabrication, each focus selects one pixel from its region, cycling through the four sub-blocks in a randomized order. This guarantees that all 25 simultaneously exposed pixels are separated by at least 10 pixels (>5 μm). Meanwhile, this sub-block strategy combined with randomized sequencing guarantees that neighboring pixels are exposed at intervals well beyond 20 ms, thereby minimizing diffusion accumulation and pixel crosstalk. In contrast, the sequential scanning shown in Supplementary Fig. S2b produces blurred boundaries and stitching artifacts, as also demonstrated in movie S1.

Optical setup for image classification

A schematic of the experimental setup is provided in Supplementary Fig. S3. The system employs a 520 nm green continuous-wave laser (Oeabt OM-12A520-3-G, 3 mW output power) as the light source. The optical configuration begins with a 4-f beam expansion system using lenses with focal lengths of 25.4 mm and 150 mm to increase the speckle diameter. A polarization modulation stage consisting of a 45° linear polarizer, followed by a phase-modulating SLM (UPOLabs HDSLM80R Plus) and a 135° linear polarizer, enables intensity modulation through polarization interference. The input images from different datasets are displayed and projected onto an SLM positioned in front of the ONN image sensor. The optical preprocessing layer employs a random phase mask for dimensionality reduction. After modulation, a second 4f system reduces the speckle diameter to match the DOE dimensions before final imaging through the DOE and focusing lens onto either a scientific CCD camera (ThorLabs Kiralux CS235CU) or alternatively a 10 × 10 photodetector array. The captured optical output is processed through a single-layer fully connected network to generate classification results. Additional experimental details can be found in the Methods section. This

configuration maintains the advantages of compact optical processing while enabling efficient data acquisition and analysis.

Training Neural network at digital backend

Supplementary Fig. S4 details the training process of the electronic neural network at the digital backend, which comprises three sequential stages: (1) Image Preprocessing: For the captured light field output image, the central region (such as 100 × 100 pixels) is first cropped and subsequently downsampled to a resolution of 10 × 10 pixels; (2) Feature Input: The downsampled image is flattened into a feature vector $\boldsymbol{x}$ and fed into a single-layer fully connected network; (3) Classification Output: Using the fashion-MNIST dataset as an example, 1000 weight parameters $\boldsymbol{W_t}$ are trained to generate the final classification result $\boldsymbol{y}$, which can be mathematically formulated as:

$$\boldsymbol{y} = softmax(\boldsymbol{b} + \boldsymbol{W_t x}) \tag{S4}$$

Here, $\boldsymbol{W_t}$ denotes the trainable weight parameters, $\boldsymbol{b}$ represents the bias term, and the $softmax$ activation function is applied to generate probabilistic classification output.

Supplementary Fig. S5 depicts the training and testing accuracy trend curves of downsampled experimental results (10 × 10 pixels) for four datasets in this study (down-sampling was performed by nearest-neighbor interpolation using OpenCV (Python): MNIST (Supplementary Fig. S5a), fashion-MNIST (Supplementary Fig. S5b), Weizmann dataset (Supplementary Fig. S5c), and flow cytometry dataset (Supplementary Fig. S5d). Supplementary Fig. S5e and Fig. S3f present the corresponding accuracy trend curves for the Fashion-MNIST and CIFAR-10 datasets under 50 × 50 pixels downsampling conditions. The optimization network training was complemented on the basis of Python 3.12 and Pytorch 2.3.1. The optimization framework employed Adam algorithm over 40 training epochs, with parameter updates driven by minimization of negative log-likelihood loss between predicted probabilities and ground-truth labels. All configurations demonstrated asymptotic convergence, as evidenced by stabilized accuracy trajectories across all datasets by terminal epoch. Computational workflows were accelerated through an Intel Core i7-13620K/NVIDIA RTX 3080 Ti hardware configuration.

Transmission efficiency of the DOEs

All fabricated devices demonstrated a transmission efficiency exceeding 70% (detailed quantitative analysis in Table S3). We apply a custom-built multi-focus TPL system described in the Methods section of the main text to generate five unique neural density configurations. These configurations are integrated into periodic nanostructures with lattice constants ranging from 500 to 1200 nm. Each configuration occupies an active area of 1 mm^2, which is constructed by assembling 50 × 50 μm^2 sub-functional units in a tiled pattern. The phase modulation architecture employs five discrete gradient levels achieved through height-variant surface relief structures (100–500 nm topography, corresponding to $\pi/5$ phase increments per 100 nm step). Material stack consists of photoresist (n = 1.520 for the wavelength of 520 nm) patterned on soda-lime glass substrates.

Quality optimized stitching for large area thin layer microstructures

To attain optimum optical performance, the 1 mm-thick glass substrate was cleaned sequentially in acetone and isopropyl alcohol (each for five minutes at ambient temperature), then dried with nitrogen to ensure reliable photoresist adhesion. The DOE layers feature deliberately varied (quasi-random) thickness profiles determined by the focal depth (z-position) of the laser relative to the substrate surface. Given the inherent stability of the optical focus during exposure, large-area stitching is executed through the six-axis nano-positioning stage.

For sub-micron (nano-scale) thicknesses spanning millimeter-scale widths, stitching-coordinate precision is critical. First, the high-precision nano-positioning stage aligns the coordinates of the glass surface and the focus-scanning system, thereby eliminating tilt errors in all degrees of freedom. Second, substrates with superior surface flatness are clamped securely to prevent local bending. These measures allow the hexapod to correct residual tip-tilt and alignment errors, resulting in accurate assembly of the microstructure (Supplementary Fig. S6a). For comparison, Supplementary Fig. S6b–S6c illustrate defects arising from conventional raster-scanning, which induce local thickness variations, missing or detached features that degrade optical-information transmission and impair subsequent neural-network computations.

Impact of image downsampling on accuracy

The size of the downsampled image directly determines the number of trainable parameters in the digital backend of the single-layer network, thereby influencing its output performance. To systematically investigate this relationship, we conduct a comprehensive analysis of how

variations in downsampling scale affect classification outcomes. As evidenced by the experimental results in Table S4, a clear positive correlation exists between the downsampling scale and classification accuracy-larger scales consistently yields improved recognition performance.

Setup for the face keypoint detection experiment

Supplementary Fig. S7a illustrates the architecture of our opto-electronic neural network for facial point detection. The system processes 96 × 96-pixel facial images through an optical encoder, followed by a lightweight digital processing backend (EfficientNet-V2) that serves as the feature extraction decoder. Notably, the same digital architecture simultaneously processes both raw SLM-displayed images (without optical sampling) and optically transformed images captured through the DOE-modulated imaging system, as shown in Supplementary Fig. S7.

Supplementary Fig. S7b shows characteristic random speckle patterns acquired by the CCD sensor, providing visual evidence that facial information becomes securely encrypted through optical scattering phenomena. This physical mechanism inherently enhances privacy protection by implementing optical encryption during data acquisition while improving adversarial robustness through light scattering's physically unclonable characteristics.

Setup for the double-layer DOE experiment and classification results for CIFAR-10 dataset

The experimental setup for testing the double-layer DOE is shown in Supplementary Fig. S8a. The optical path design principle is consistent with that described in the previous section. The difference is that the single-layer random phase mask design has been replaced by a double-layer DOE design. A 3D-printed holder is used to secure the DOE samples, maintaining a spacing of 3 mm between the two layers, as illustrated in Supplementary Fig. S8b. The corresponding training and testing accuracy curves for the downsampled CIFAR-10 dataset (50 × 50 pixels) are provided in Supplementary Fig. S8c. The results indicate that, compared with the single-layer configuration, the classification accuracy achieved with the double-layer DOE has improved to 95%.

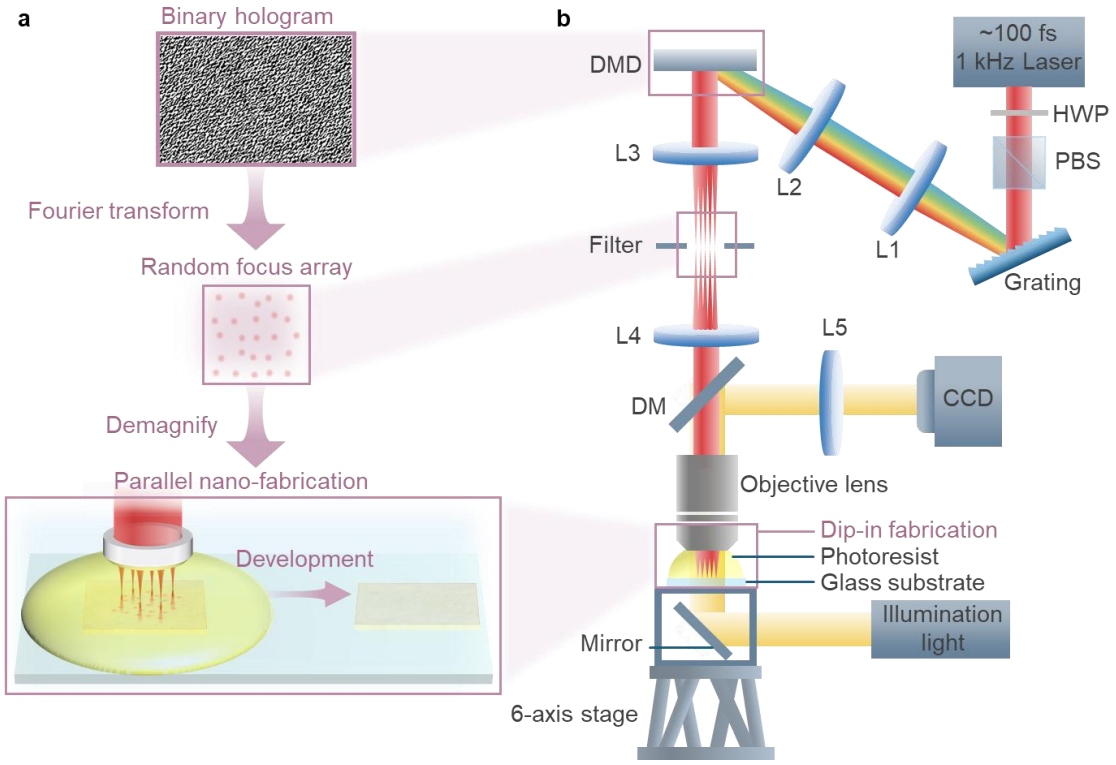

Figure S1. Schematic of the randomized multi-focus TPL nanofabrication system. **a,** Working principle: During fabrication, the DMD sequentially projects random-access multi-focus binary holograms; the focus array is generated on the Fourier plane of DMD; and the focus array is demagnified and exposure frame-by-frame under the objective lens in a dip-in manner; after development, the solidified DOE structure remains adhered to the glass surface while the unexposed photoresist is removed. **b,** Optical configuration: HWP, half-wave plate; PBS, polarizing beam splitter; DM, dichroic mirror; L1–L5: lenses ($f_{L1}, f_{L2}, f_{L3}, f_{L4}, f_{L5}$ = 225, 250, 150, 200, and 125 mm); M1: Mirror.

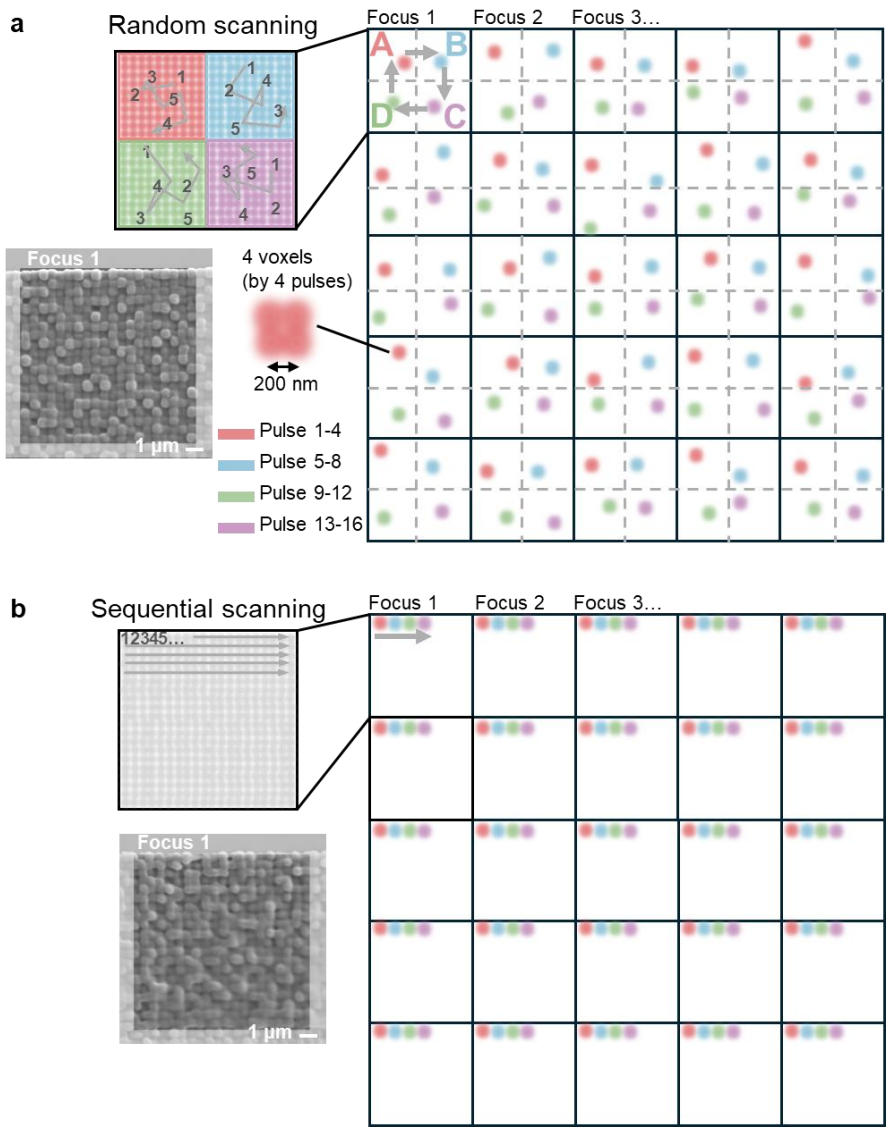

Figure S2. Randomized sub-block scanning strategy. **a,** Random scanning and the SEM image result. **b,** Sequential scanning and the SEM image result.

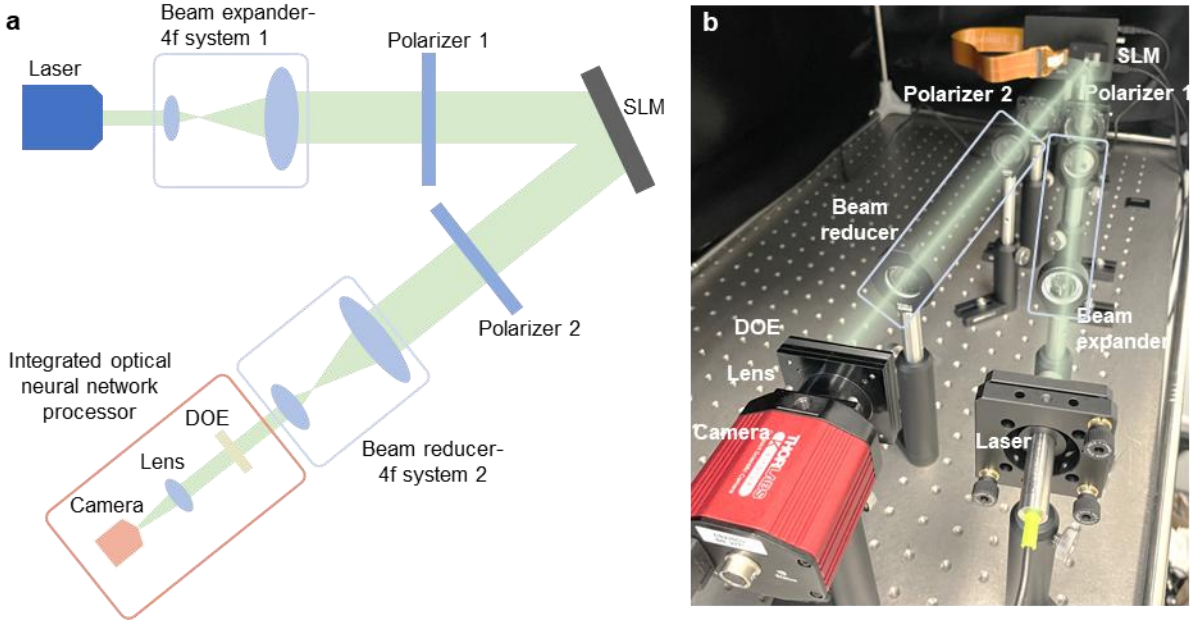

Figure S3. Optical setup for image classification. **a,** Schematic diagram of the optical setup. **b**, Photo of the optical setup.

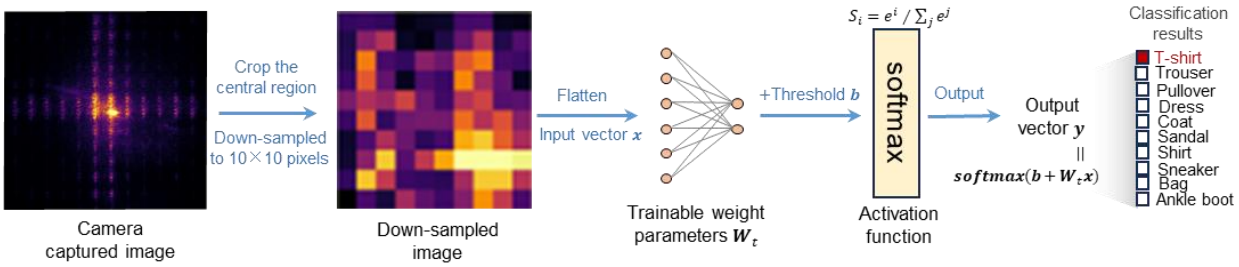

Figure S4. Flowchart of the electronic neural network training process.

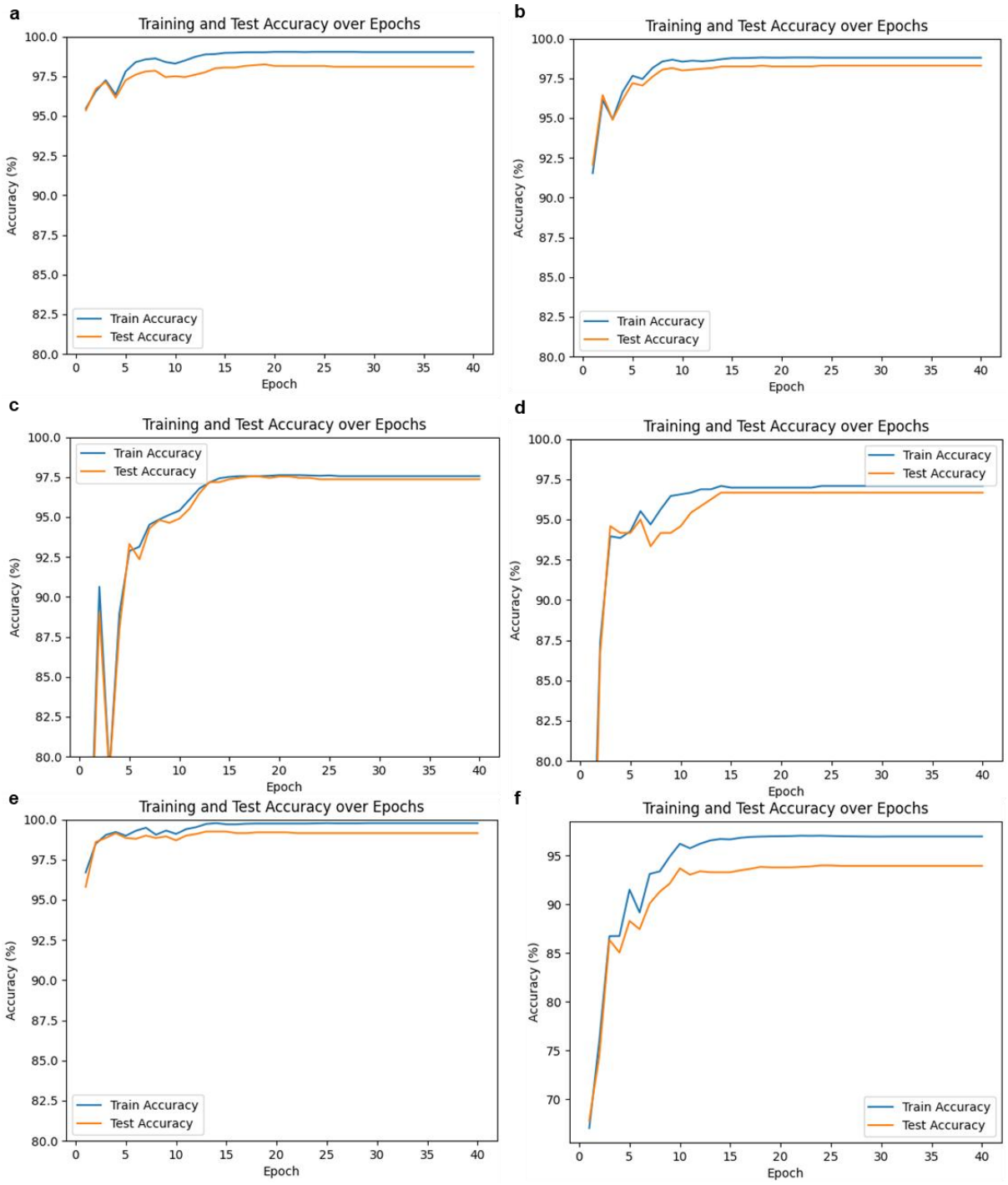

Figure S5. Training and testing accuracy trend curves of downsampled experimental results (10 × 10 pixels) for four datasets: **a**, MNIST, **b**, Fashion-MNIST, **c,** Weizmann dataset, and **d,** flow cytometry dataset. **e** and **f** present the corresponding accuracy trend curves for the Fashion-MNIST and CIFAR-10 datasets under the 50 × 50 pixels downsampling conditions.

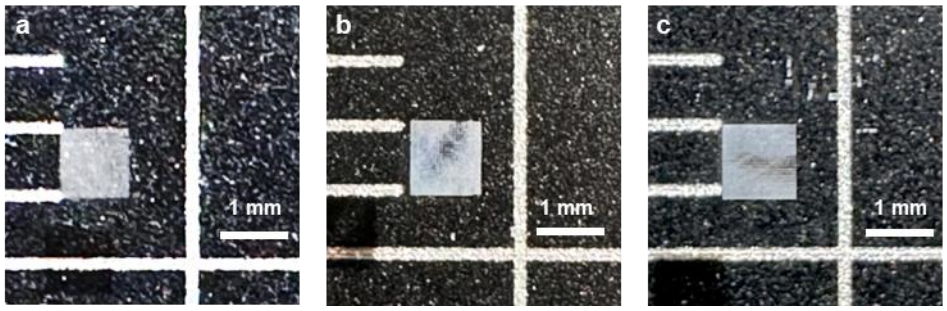

Figure S6. Optical images of ONNs with different stitching qualities. **a**, Accurately stitched ONN. **b and c,** Badly stitched ONNs.

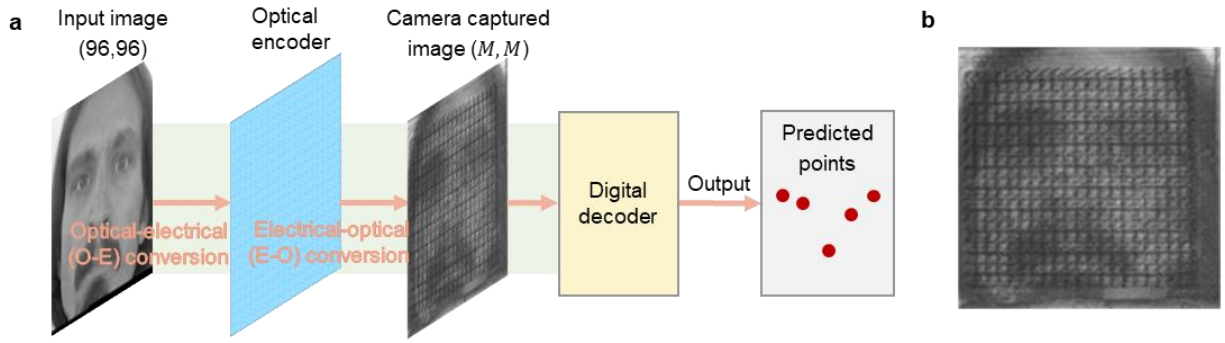

Figure S7. a, Optoelectronic neural network for facial point detection. **b,** Random speckle patterns acquired by the CCD sensor.

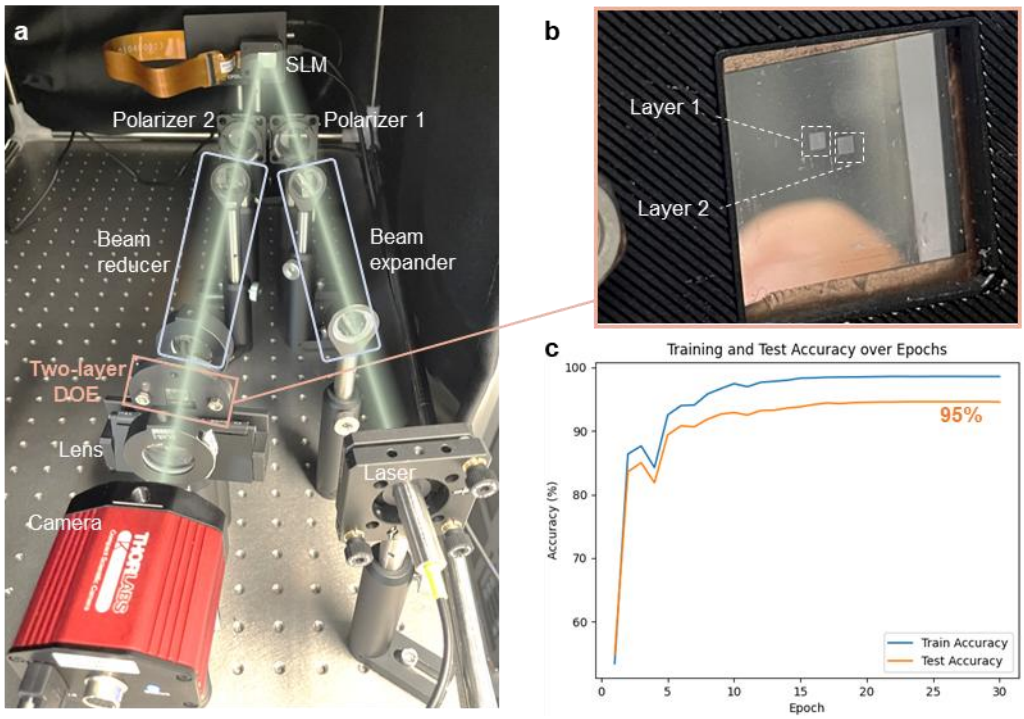

Figure S8. a, Photos of the optical setup. **b,** Double-layer diffractive optical element (DOE). **c,** Training and testing accuracy trend curves of the downsampled experimental results (50 × 50 pixels) for the CIFAR-10 dataset.

Table S1. Device comparison with the reported works including both 3D printing, photolithography and EBL (such as neuron number, area, density, fabrication time and speed).

Ref.	Device	Neuron number	Neuron Density (million/mm²)	Fabrication time/ area	Speed (thousand neurons/min)
[8]	3D-printed DOE	0.04 million	6.25×10^{-6}	6.5 min/ 6.4 mm^2	6.15
[12]	3D-printed DOE	0.01 million	6	1–2 min/ 2.5×10^{-3} mm^2	5
[11]	DOE based on photolithography	1 million	6.25×10^{-2}	2.5–4 h/ 16 mm^2	6.67
[18]	3D-printed DOE	0.0144 million	6.25×10^{-6}	4.7 min/ 2.304 mm^2	3.06
[44]	Metasurface	0.0784 million	6.25	1.96 h/ 1.2544×10^{-2} mm^2	0.67
[45]	Metasurface	0.339 million	4.938	10.72 h/ 6.865×10^{-2} mm^2	0.72
[43]	DOE based on photolithography	1 million	4×10^{-2}	2.5–4 h/ 2.5 mm^2	6.67
Our work	**3D-printed DOE**	**4 million**	**4**	**15 min / 1 mm^2**	**266.67**

Table S2. Performance comparison with previously reported works on free-space ONNs.

Dataset	Ref.	Device	Neuron number	Wavelength	Accuracy	Training for optical neurons
MNIST/ Fashion-MNIST	[8]	3D-printed DOE	0.2 million	750 μm	93.39% (MNIST) / 86.60% (Fashion)	Need
	[12]	3D-printed DOE	0.01 million	785 nm	86.67% (MNIST)	Need
	[11]	DOE based on photolithography	1 million	632.8 nm	84% (MNIST)	Need
	[18]	3D-printed DOE	0.0144 million	600–1200 μm	87.74 ± 1.12% (MNIST)	Need
	[28]	SLM	/	532 nm	92% (MNIST)	No need (Digital readout parameters:40960)
	[30]	Disordered polycrystalline slab	/	800 nm	96.5% (MNIST) /87.6% (Fashion)	No need (Digital readout parameters: 35000)
	[32]	Random aberrations and defects	/	532 nm	87%	No need (Digital readout parameters:40000)
	Our work	**3D-printed DOE**	**4 million**	**520 nm**	**98% (MNIST) / 98% (Fashion)**	**No need (Digital readout parameters: 1000)**
CIFAR10	[43]	DOE based on photolithography	1 million	532 nm	44.4%	Need
	[30]	Disordered polycrystalline slab	/	800 nm	52%	No need (Digital readout parameters:35000)
	Our work	**3D-printed DOE**	**4 million**	**520 nm**	**94% (single-layer) /95% (double-layer)**	**No need (Digital readout parameters: 25000)**

Table S3. Transmission efficiency of different 3D-printed layers with unit periods of 500 nm, 600 nm, 700 nm, 800 nm, and 1200 nm.

Parameters	Transmission efficiency (%)
p = 500 nm	82.1
p = 600 nm	95.1
p = 700 nm	91.9
p = 800 nm	70.2
p = 1200 nm	75.2

Table S4. Impact of image downsampling on accuracy on multiple datasets, including MNIST, Fashion-MNIST, and CIFAR-10.

Dataset	Downsampled image scale	Accuracy (%)
MNIST	10 × 10	98
	50 × 50	99.25
	90 × 90	99.5
Fashion-MNIST	10 × 10	98
	50 × 50	99.15
	90 × 90	99.5
CIFAR10	10 × 10	80
	50 × 50	94
	90 × 90	95

Movie S1.

The movie illustrates the trajectory planning of the randomized multi-focus TPL nanofabrication strategy, compared with the sequential scanning method. In the first stage, a 100× slowed-down simulation demonstrates the exposure of multi-focus frames (frames 1–40) for trajectory planning. Blue dots indicate the exposure positions of the 25 foci, while the orange dots mark the polymerized pixels. Each pixel is formed by four exposures with a 200 nm spacing, thus requiring four frames to complete. In the second stage, an *in situ* capture shows the 1.6 s printing process of a 50 μm scanning area recorded by a CCD (as denoted in Supplementary Fig. S1b). The polymerized pixels are clearly visible, while the moving blue light array corresponds to fluorescence emission during the polymerization process. In the final stage, an SEM image covering a 10 μm area fabricated with a single focus is shown. Pixels generated with the randomized scanning method are well defined, whereas those produced by sequential scanning appear merged, with visible boundaries between regions scanned by different foci.