

Photonic neuromorphic learning via generalized in situ physical gradient descent

In the format provided by the
authors and unedited

Supplementary Information for **Photonic neuromorphic learning via
generalized in situ physical gradient descent**

Supplementary Note 1. Chip calibration.....	2
Supplementary Note 2. Scattering layers for convolution	3
Supplementary Note 3. Pseudocode of the in situ physical gradient descent.....	3
Supplementary Note 4. The incorporation of nonlinearity in INSPIRE	6
Supplementary Note 5. Scalability of the INSPIRE array	6
Supplementary Note 6. Full designed matrix entries of the 4×4 matrix in Fig. 2..	9
Supplementary Note 7. Training in a wavelength-multiplexed system.	10
Supplementary Figures	11
Supplementary Table 1. Comparison with existing photonic training methods.	21

Supplementary Note 1. Chip calibration

Electrical crosstalk. The controlling voltage between different ports depends on several neighboring ports. Because of the non-negligible voltage drop on wires, the control of one port may modify the voltage of several ports. Therefore, we first calibrated the electrical crosstalk matrix, obtaining the mapping relationship between the voltage across the electrical tuning components and the output voltage of the signal source. For each of the 82 voltage driver pins, a single pin was varied in voltage from 0 to 10.0 V while each of the other related pins was probed. The probed voltage difference (negative value) was saved into the corresponding element of the 82×82 matrix, which is shown in Fig. S2. The target voltage vector could be multiplied by the crosstalk matrix to get the actual input value. Given the target voltage matrix, the driving voltages can be acquired by multiplying the inverse of the crosstalk matrix. To further minimize the thermal crosstalk, pulsed signals were used to drive the modulators.

Control voltage mapping. Following previous methods, we first calibrated the electrical crosstalk matrix. Subsequently, all voltages applied to the electrical tuning components were mapped to the signal source output via the matrix to control the scan calibration. The principle of scan calibration is determined by the electro-optical modulation characteristics of the thermal phase shifter. Thus, the observed intensity follows the relationship with the input voltage described as follows:

$$I = a \times \sin(b \times x^2 + c) + d \quad (\text{S1})$$

where I denotes intensity, x denotes input, and a, b, c, d are the trainable parameters. During the scan, the voltage range was set from 0 to 10 V (with phase shifters having a scanning upper limit slightly above 10 V), and 150 points were scanned at equal intervals. This ensured that the modulation curve shows at least one complete cycle, which is then truncated at the end to capture exactly one full cycle. Subsequently, based on the above model, the phase corresponding to each point was fitted, determining the relationship between the relative phase of the two arms and the input voltage. The 16-input modulators, error field modulators, the four calibration modulators, and 4×4 mesh were calibrated sequentially. The calibration of the error field modulators was performed using the reverse-input configuration.

Phase retardation. To faithfully measure both the hologram of the data field and that of the error field, it is necessary to align the reference beam to the same phase front. However, due to the varying on-chip propagation distances, the reference points of the 36 STR units exhibit disrupted phases. In the experiment, prior to implementing the in situ design, we calibrated the phase profiles by constructing an on-axis holographic setup. As illustrated in Fig. S4, the laser source was split into two beams: one was coupled onto the INSPIRE chip under test (DUT), while the other was collimated into free space. The on-chip beam was directed into the reference port and emitted outward, after which the output light was relayed through a 4f system with an expansion

coefficient of two. The plane wave from the collimator was then combined with the emission from the chip using a beam splitter, producing interference patterns on the infrared camera. To calibrate the phase retardation, the phase of the on-chip reference beam was swept over a full π range through thermal phase tuning. The resulting interference data were fitted to Eq. (S1), yielding the relative phase c which characterizes the phase front.

Supplementary Note 2. Scattering layers for convolution

Convolution is extensively utilized in applications such as signal analysis and image processing. Recently, the integration of convolutional operations with neural network methods, known as Convolutional Neural Networks (CNNs), has gained popularity. Signals are first convolved with pre-defined kernels to extract features before being fed into fully connected layers. CNNs are applied in diverse fields, including image classification, autonomous navigation, and audio spectrum analysis. Several research efforts have explored the use of optical computing for analogue convolution. Compared to electrical methods, these techniques offer significant improvements in computational efficiency, speed, and parallelism.

Here, we provide a simulated demonstration of the efficient convolution operation on input signals achieved by diffractive structures. In this example, the input signal consists of 25 pixels, and the diffractive structure performs a 16-dimensional convolution, resulting in a 10-pixel output. The signal is encoded onto the amplitude of the input light, propagates through a slab waveguide and three layers of diffractive masks, and the intensity is finally captured by ten output waveguides in the detection layer. A comparison of optical and digital convolution shows that the results of optical convolution (after applying a fixed compensation factor) match those of digital convolution. The error distribution between digital and optical convolution results, shown in Supplementary Figure 5, indicates an error of 1.61×10^{-2} (four orders of magnitude smaller than the average intensity). This demonstrates that the optical diffraction convolution layer can extract signal features similarly to digital convolution, while offering higher processing speed and energy efficiency.

Supplementary Note 3. Pseudocode of the in situ physical gradient descent

We utilize the in situ physical gradient descent function to achieve in situ photonic training. This function includes sub-functions for data propagation and error propagation to retrieve the fields of STR units for the calculation of the on-chip gradients. To measure the respective fields, the phase and amplitude were recorded through the on-chip hologram function. The specific workflow of these functions is as follows.

Algorithm In situ physical gradient descent

```
1 mainFunction ForwardData ( $x, T, M$ )
2   Input data:  $x \in \mathbb{C}^{S \times 1}$ ; target:  $T \in \mathbb{C}^{K \times 1}$ ; Loss function:  $L$ ;
3     reconfigurable modulation parameters:  $M = [M_1, M_2, \dots, M_N]$  ( $M_k =$ 
4        $\exp(2\pi j n_R d / \lambda), n_R \in R, k = 1, 2, \dots, N$ )
5   Output output gradient:  $g \in \mathbb{R}^{1 \times N}$ 
6    $x = x / \|x\|$ 
7    $E_{data}, E_{out} = \text{SystemData}(x)$ 
8    $E_{loss} = \text{GradFunction}(T, E_{out}, L)$ 
9    $E_{err} = \text{SystemError}(E_{loss})$ 
10   $g \leftarrow \text{Real}(j E_{data} E_{err})$ 
11  return  $g$ 
12 end function
13 subFunction1 SystemData( $x$ )
14   Input data:  $x \in \mathbb{C}^{S \times 1}$ 
15   Output output field:  $E_{out} \in \mathbb{C}^{K \times 1}$ ,  $E_{data} \in \mathbb{C}^{N \times 1}$ 
16   Feed  $x$  to the input ports
17    $E_{data}, E_{out} = \text{Hologram}(N + K)$ 
18   return  $E_{data}, E_{out}$ 
19 end function
20 subFunction2 SystemError( $x$ )
21   Input data:  $x \in \mathbb{C}^{K \times 1}$ 
22   Output output field:  $E_{err} \in \mathbb{C}^{N \times 1}$ 
23   Feed  $x$  to the error ports
24    $E_{err} = \text{Hologram}(N)$ 
25   return  $E_{err}$ 
26 end function
27 subFunction3 GradFunction ( $T, E_{out}, L$ )
28   Input target:  $T \in \mathbb{C}^{K \times 1}$ ; Output field:  $E_{out} \in \mathbb{C}^{K \times 1}$ ; Loss function:  $L$ 
29   Output loss field:  $E_{loss} \in \mathbb{C}^{K \times 1}$ 
30    $E_{loss} = \partial L(E_{out}, T) / \partial E_{out}$ 
31   return  $E_{loss}$ 
32 end function
33 subFunction4 Hologram( $x$ )
34   Input number of ports to retrieve:  $x \in N$ 
35   Output output field:  $E \in \mathbb{C}^{x \times 1}$ 
36    $I_0 \leftarrow$  turn reference beam off and measure  $x$  ports
37    $I_1 \leftarrow$  turn other beam off and measure  $x$  ports
38    $I_2 \leftarrow$  turn reference beam on and measure  $x$  ports
39    $I_3 \leftarrow$  shift reference beam by  $\pi/2$  beam on and measure  $x$  ports
40    $A = \sqrt{I_0}$ 
41    $A_{ref} = \sqrt{I_1}$ 
```

```

42      $\phi = \arccos ((I_2 - I_0 - I_1)/(2\sqrt{I_0 I_1}))$ 
43     if  $I_3 \neq |A_{ref} + jAe^{j\phi}|^2$ 
44          $\phi = -\arccos ((I_2 - I_0 - I_1)/(2\sqrt{I_0 I_1}))$ 
45          $E = Ae^{j\phi}$ 
46         return  $E$ 
47 end function
48

```

```

49 Notations:
50  $S$ : number of input ports
51  $K$ : number of output ports
52  $N$ : number of STR units

```

Supplementary Note 4. The incorporation of nonlinearity in INSPIRE

INSPIRE can be extended to multilayer nonlinear PNNs. As shown in Fig. S7a, under the general form of nonlinear activation, $x^k = f(y^{k-1})$, the error propagation can be formulated as $\delta(y^{k-1}) = f'(y^{k-1})\delta(x^k)$, which indicates that the error field is modified by the adjoint Jacobian of the nonlinear activation, i.e., the error field entering the preceding photonic layer should be multiplied by $f'(y^{k-1})$ or its complex-valued generalization. For pointwise nonlinearities, this operation can be implemented by complex modulation of the error-propagating light before it is injected into the preceding layer. In this way, the STR unit still provides the local field needed for physical gradient calculation, while the nonlinear derivative is incorporated through an additional modulation step in the error path. The simulation results on the MNIST and F-MNIST datasets are compared in Fig. S7b. We introduce an intensity-induced nonlinearity $f(y^{k-1}) = ye^{jq|y^{k-1}|^2}$, where the coefficient q was set to 0.5 in the simulation. The incorporation of nonlinearity improved the accuracies by 4.27% and 14.56% in the F-MNIST and MNIST datasets, respectively. The proposed extension preserves the shared data/error-field formalism of INSPIRE and may offer a compact route to integrating nonlinear activation derivatives into the physical backpropagation path.

Supplementary Note 5. Scalability of the INSPIRE array

Electrical control. In the current prototype, electrical access is implemented through wire bonding, whose routing complexity increases with the number of STR elements. This issue can be mitigated in future scaled systems through three-dimensional heterogeneous integration or co-packaging of the photonic chip with the electronic control/readout chip. Such integration would enable point-to-point addressing of detectors and modulators while substantially reducing routing complexity. As the array size increases, the number of STR units increases accordingly. However, because each STR unit modulates the broadcast optical signals locally and the units do not interfere with one another during calibration, their modulation responses can, in principle, be

calibrated in parallel. This parallel calibration capability helps mitigate the increase in calibration time for larger-scale arrays.

Optical signals. Here we provide an insertion-loss and signal-to-noise ratio (SNR) analysis as a function of the number of STR units. The analysis is decomposed into two parts, the insertion loss of signal light and reference light, and the signal noise.

The total insertion loss is decomposed into three contributions: STR splitting/tapping loss, propagation loss, and waveguide-crossing loss, denoted as l_s , l_p , and l_x , respectively. First, for a $\sqrt{N} \times \sqrt{N}$ array on a silicon-on-insulator platform, the multi-stage splitting loss is $10\log_{10}(\sqrt{N})$ and the splitting tree itself has a loss of $Ds \cdot \log_2 N$, where Ds is the splitting loss, normally less than 0.5 dB. Each optical path of the array includes two splitter losses, one at the modulator and one at the final detector. Assuming 50/50 directional couplers, as used in this work, each splitter introduces a 3-dB loss; therefore, the total splitting loss is $l_s = -3 \text{ dB} \times 2 - 10\log_{10}(\sqrt{N}) - Ds \cdot \log_2 N = -6 \text{ dB} - 10\log_{10}(\sqrt{N}) - Ds \cdot \log_2 N$. Second, for the propagation loss, the array pitch is set to $p = 700 \text{ }\mu\text{m}$, consistent with our device design. The longest path length in a $\sqrt{N} \times \sqrt{N}$ array is approximately $2\sqrt{N}p$. For the STR loop, the propagation distance is $300 \text{ }\mu\text{m}$. Considering a typical SOI waveguide propagation loss of -2 dB/cm , the propagation loss is estimated as $l_p = -2 \text{ dB/cm} \times (2\sqrt{N}p + 300 \text{ }\mu\text{m}) = (-0.28\sqrt{N} - 0.06) \text{ dB}$. Third, each detection path contains at most $2(\sqrt{N} - 1)$ waveguide crossings, including $(\sqrt{N} - 1)$ crossings with the reference-light routing and $(\sqrt{N} - 1)$ crossings with the data-light tapping paths. Assuming a crossing loss of -0.1 dB per crossing, the total crossing loss is $l_x = 2(\sqrt{N} - 1) \times -0.1 \text{ dB} = -0.2(\sqrt{N} - 1) \text{ dB}$. Therefore, the total insertion loss is given by

$$l = l_s + l_p + l_x = \left(-6 \text{ dB} - 10\log_{10}(\sqrt{N}) - Ds \cdot \log_2 N - 0.28\sqrt{N} - 0.06 - 0.2(\sqrt{N} - 1) \right) \text{ dB}. \quad (\text{S2})$$

Simultaneously, the reference fan-out loss consists of the insertion loss $10\log_{10}(\sqrt{N})$ and splitting loss $Ds \cdot \log_2 N$, such that the reference fan-out loss

$$l_{\text{fan}} = -10\log_{10}(\sqrt{N}) - Ds \cdot \log_2 N. \quad (\text{S3})$$

Noise level.

The mutually independent TIA, shot-noise, and RIN terms are written as:

$$\sigma_{\text{tot},i}^2 = \sigma_{\text{TIA}}^2 + \sigma_{\text{shot},i}^2 + \sigma_{\text{RIN},i}^2. \quad (\text{S4})$$

Let i_{TIA} denote the total input-referred current-noise density of one detector-TIA channel, including the amplifier and feedback-network contributions. We assume that this density is approximately white over the equivalent noise bandwidth B , the rms electronic noise is $\sigma_{\text{TIA}} = i_{\text{TIA}}\sqrt{B}$. In this calculation, we assume i_{TIA} has a typical value of 0.9 pA/ $\sqrt{\text{Hz}}$. The shot noise can be calculated as $\sigma_{\text{shot},i} = \sqrt{2qI_iB}$, where q is the elementary charge and I_i is the total mean direct current in the form of $I_i = \mathcal{R}(P_{s,i} + P_{r,i}) + I_{d,i}$. $I_{d,i}$ represents the dark current value, normally around 30 nA.

The relative -intensity noise term can be written as $\sigma_{\text{RIN},i} = \mathcal{R}(P_{s,i} + P_{r,i})\sqrt{\text{RIN}B}$, where the RIN value is on the order of 10^{-14} Hz^{-1} .

The total noise variance becomes $\sigma_{\text{tot},i}^2 = (i_{\text{TIA}}\sqrt{B})^2 + 2q(\mathcal{R}(P_{s,i} + P_{r,i}) + I_{d,i})B + (\mathcal{R}(P_{s,i} + P_{r,i})\sqrt{\text{RIN}B})^2$.

To summarize the above discussion, the detector powers are $P_{s,i} = P_{s,0}10^{l/10}$, $P_{r,i} = P_{r,0}10^{l_{\text{fan}}/10}$. The amplitude of the coherent interference current is $i_{\text{hol},i} = 2\mathcal{R}V_i\sqrt{P_{s,i}P_{r,i}}$, and the final SNR term becomes

$$\text{SNR}_i = 10\log_{10}\left(\frac{i_{\text{hol},i}}{\sigma_{\text{tot},i}}\right), \quad (\text{S5})$$

At the bandwidth of 100 kHz and responsivity $\mathcal{R}$ of 1 A/W, the resulting SNR_i and the insertion loss l are shown in Fig. S9.

Supplementary Note 6. Full designed matrix entries of the 4×4 matrix in Fig. 2

The target entries were sampled as: $[-20.3290-45.6808i, -42.7508+25.9300i, 49.1347+9.2617i, 7.9240+49.3681i], [-49.6697-5.7381i, -14.2726-47.9196i, 33.0553+37.5146i, -14.6768-47.7974i], [-28.5486-41.0484i, 33.7194+36.9188i, -47.3822-15.9664i, -12.3026+48.4628i], [-49.9598+2.0039i, 49.3823+7.8352i, 26.5950+42.3404i, -38.4873+31.9176i]]$.

Supplementary Note 7. Training in a wavelength-multiplexed system.

The INSPIRE method can be generalized to wavelength-division-multiplexed (WDM) systems. Supplementary Figure 12 illustrates INSPIRE training in a network comprising M cascaded units and N input wavelengths.

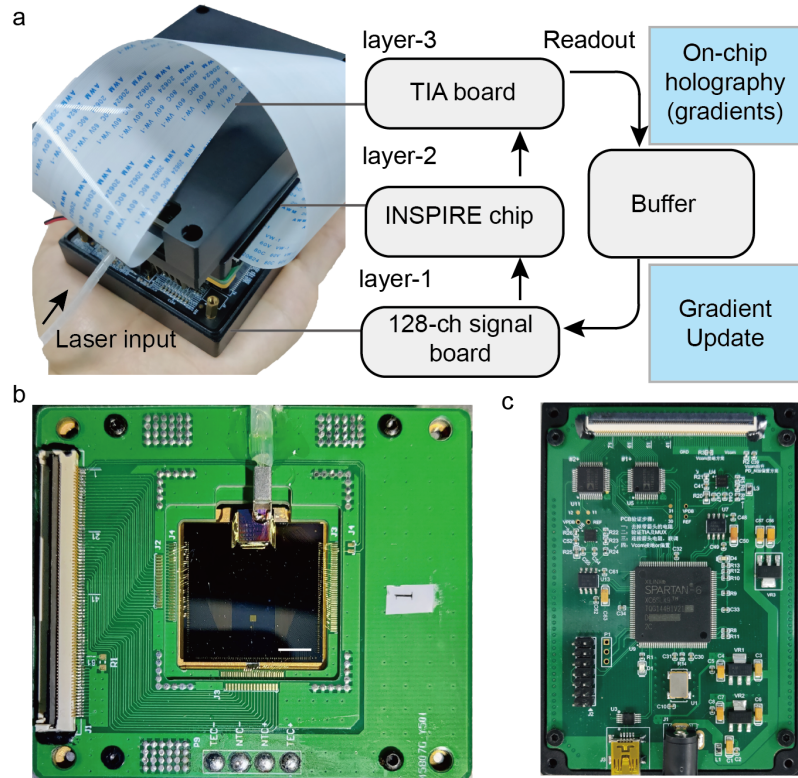
For the linear system shown in Supplementary Figure 12a, the inputs at the N wavelengths, x^{λ_i} , are sequentially injected into the neural network. For each wavelength λ_i , the data-propagation field $E_{out(j)}^{\lambda_i}$ is measured and recorded at each INSPIRE unit j , where i and j index the wavelength channels and INSPIRE units, respectively. The corresponding output y^{λ_i} is also recorded. The outputs from all wavelength channels are then used to evaluate the loss L and calculate the output error fields $E_{err}^{\lambda_i}$.

Each error field $E_{err}^{\lambda_i}$ is subsequently injected into the system, and the resulting error-propagation field $E_{err(j)}^{\lambda_i}$ is measured at every INSPIRE unit. The gradient for unit j is obtained by averaging the wavelength-resolved gradients $g_j = \frac{1}{N} \sum_{i=1}^N g_j^{\lambda_i}$, where $g_j^{\lambda_i} = \text{real}(jE_{out(j)}^{\lambda_i}E_{err(j)}^{\lambda_i})$ and $j = \sqrt{-1}$.

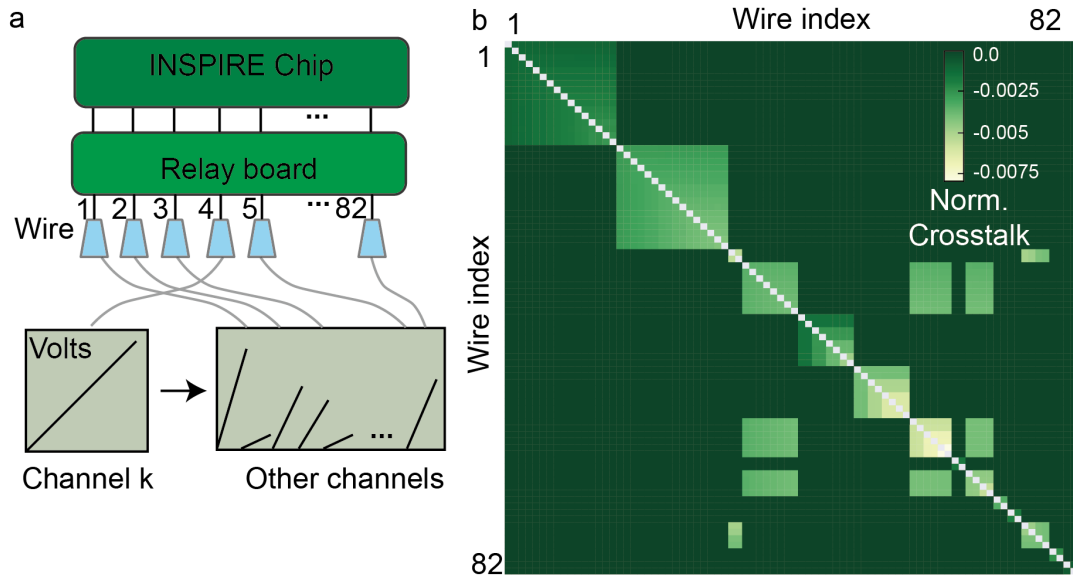
Supplementary Figure 12b illustrates the corresponding training procedure when a nonlinear activation is incorporated. The upper panel shows the propagation paths through the nonlinear unit. By controlling the phase shifter, the forward-propagating data field is routed through the nonlinear activation, whereas the backward-propagating error field passes through a modulator programmed with the derivative of the nonlinear function. During training, the wavelength-dependent inputs are first loaded sequentially and are then injected simultaneously to evaluate the nonlinear activation. The resulting activated outputs are used to calculate the error fields, which are subsequently injected into the system for error propagation.

In this illustration, MZI-based INSPIRE units are employed. However, other unit architectures, such as microring resonator (MRR) modulators, may also be feasible. Owing to their resonant nature, MRR-based implementations are more sensitive to temperature fluctuations and therefore require more precise and dedicated thermal stabilization.

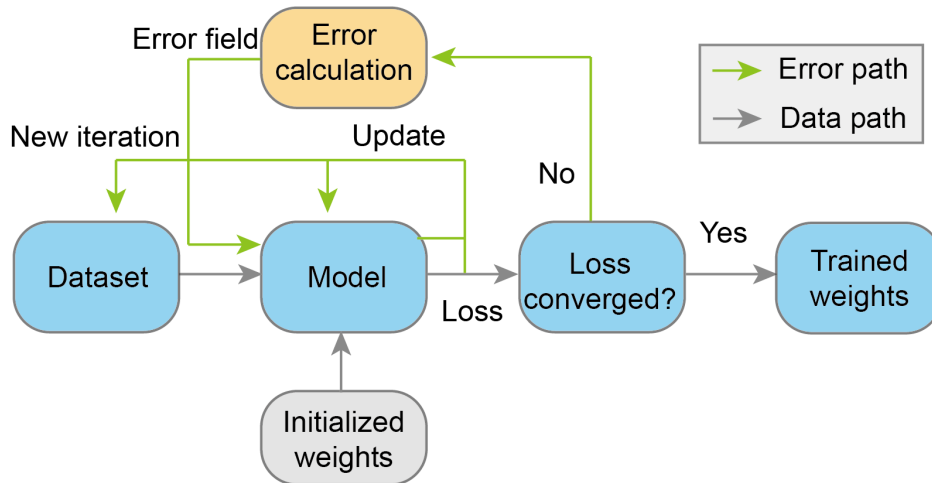
Supplementary Figures



Supplementary Figure 1. Experimental Setup. **a**, Hand-held INSPIRE system. The system consists of three functional layers. At its core lies the INSPIRE chip (layer-2), positioned between a signal source board (layer-1) and a transimpedance amplifier (TIA) board (layer-3). The signal source board provides the control signals that drive the STR units on the INSPIRE chip, while the TIA board detects the output hologram and converts it into electrical readouts. During system operation, the detected hologram is also used for gradient calculation and parameter updates in the design process. **b**, Packaged INSPIRE chip (layer-2). Optical excitation is provided by a solid-state laser (Koheras BASIK, NKT Photonics) operating at 1550 nm, which couples light into the chip through a single channel of a fiber array. The electronic pads of the chip are connected to a relayed interposer, which is wire-bonded to the printed circuit board, thereby enabling seamless integration with the overall system. **c**, TIA circuits (layer-3). The TIA board comprises 40 channels, each designed to convert photocurrent signals into voltages, with a transimpedance gain of 100 k Ω .

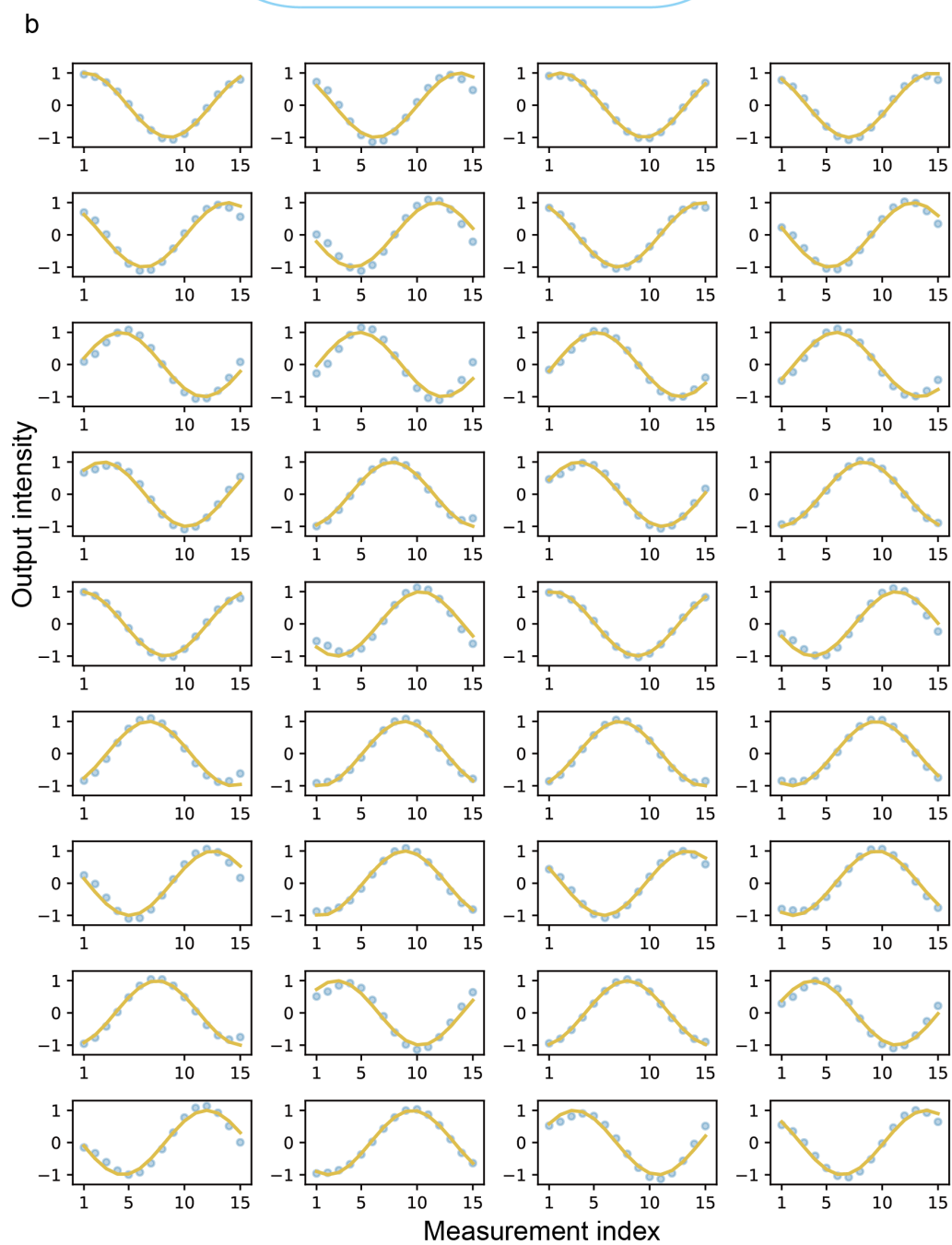
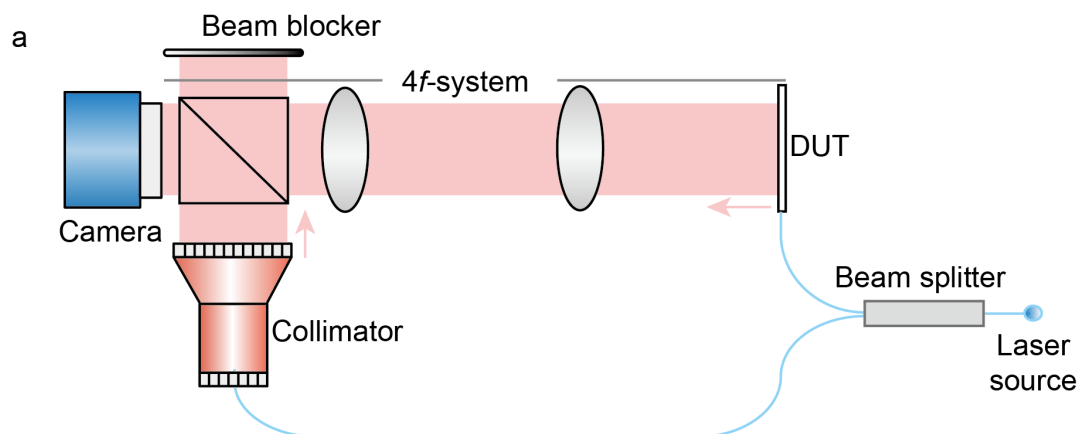


Supplementary Figure 2. Calibrated electrical crosstalk matrix. **a**, The crosstalk matrix was obtained by relaying the output from the INSPIRE chip to a relay board, where the voltage values of each wire can be collectively measured. Specifically, we applied a voltage signal to a specific wire and measured the induced voltage signals on all other wires using a multimeter. Each measured signal was normalized to the applied input signal. **b**, A total of 82 wires were calibrated in this manner. By construction, all diagonal entries of the crosstalk matrix equal one. During signal loading, the target matrix is multiplied by the inverse of the crosstalk matrix before being configured on the chip, thereby compensating for electrical interference.

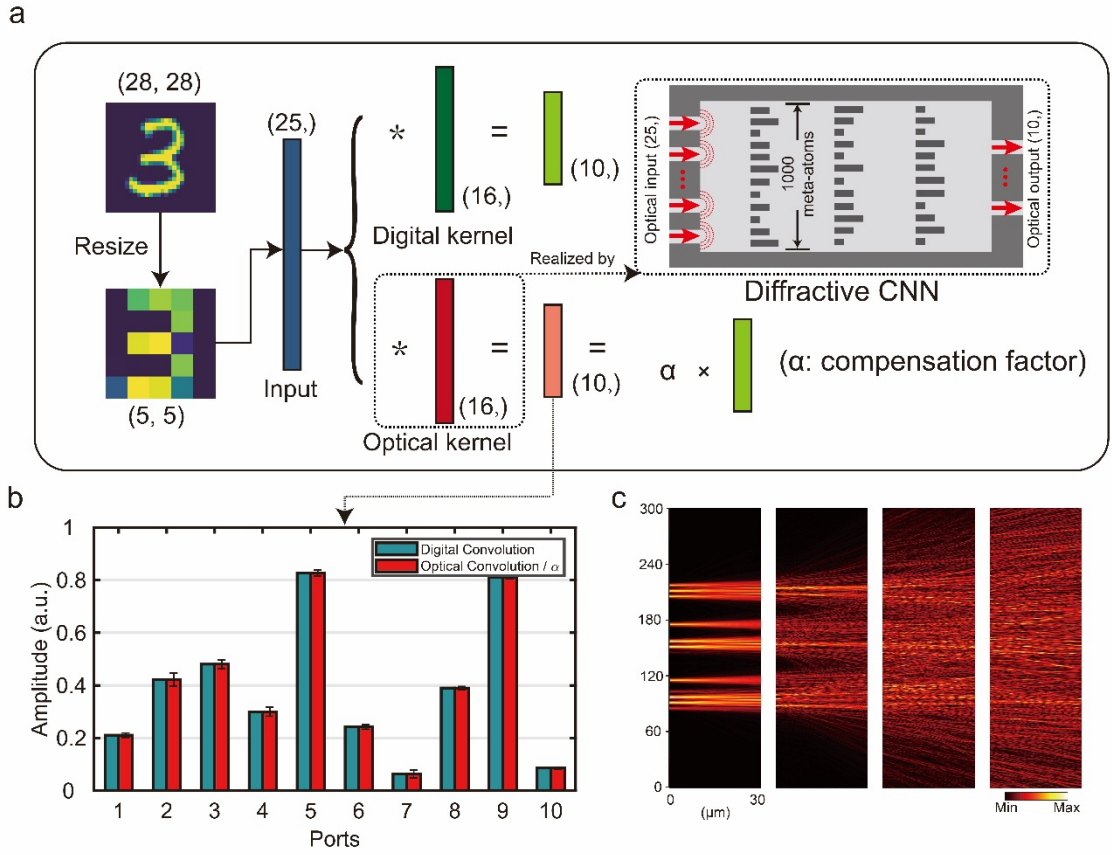


Supplementary Figure 3. Block diagram of the in situ photonic training. After initialization, the training process begins by feeding samples from the dataset into the model. In each iteration, the loss values are first evaluated, and the hologram of the data propagation is recorded. If the loss values converge, that is, if they no longer decrease over the most recent training iterations, the training terminates and final design weights are saved. Otherwise, the training process proceeds into the error propagation stage

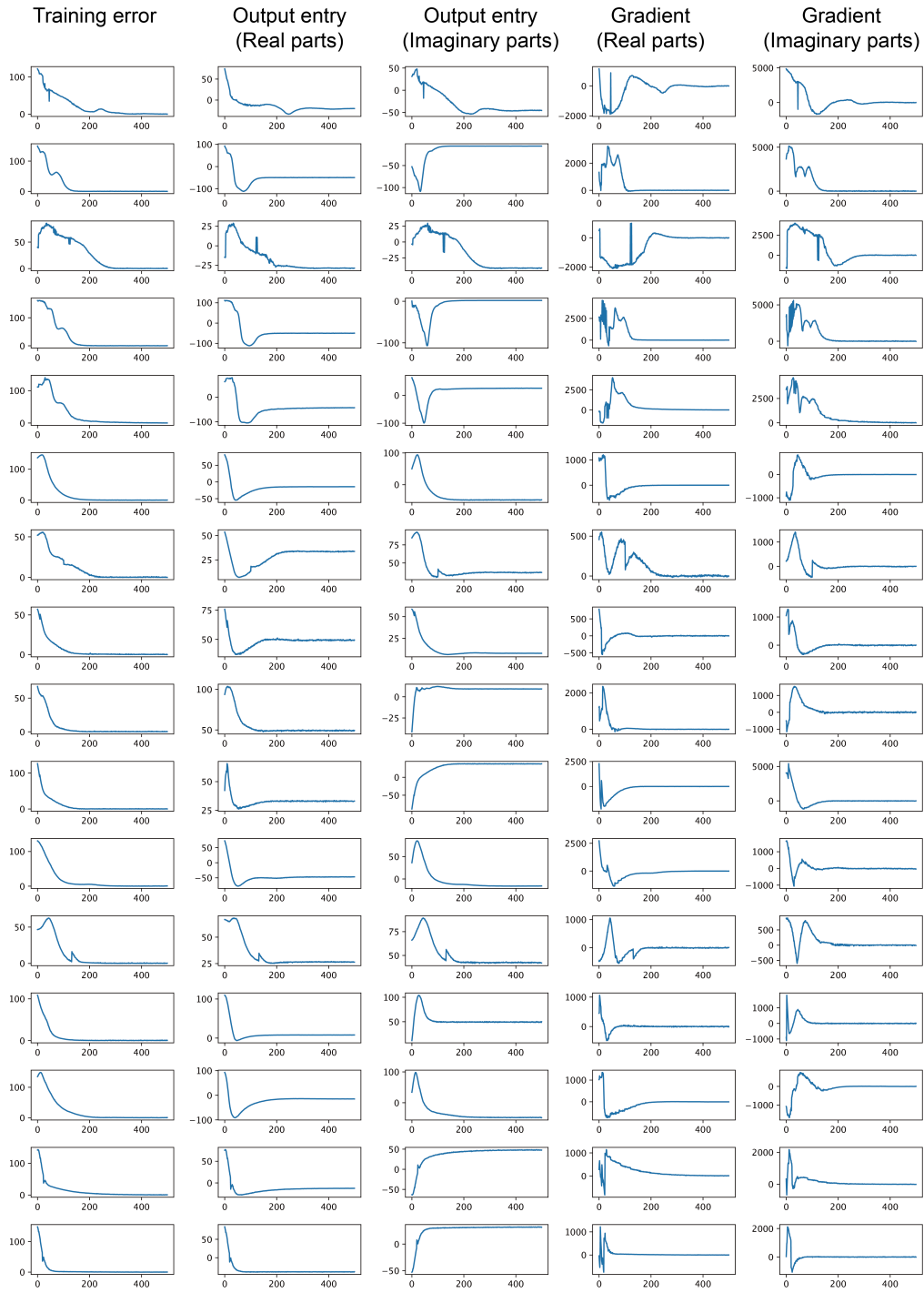
(indicated by green lines). In this stage, the error fields are computed by differentiating the loss function with respect to the output fields. The hologram of the error propagation is also recorded. Gradients are then calculated using the holograms measured in both the data and error propagation stages and subsequently used to update the model weights. After the update, a new iteration begins, and the process repeats until the loss converges.



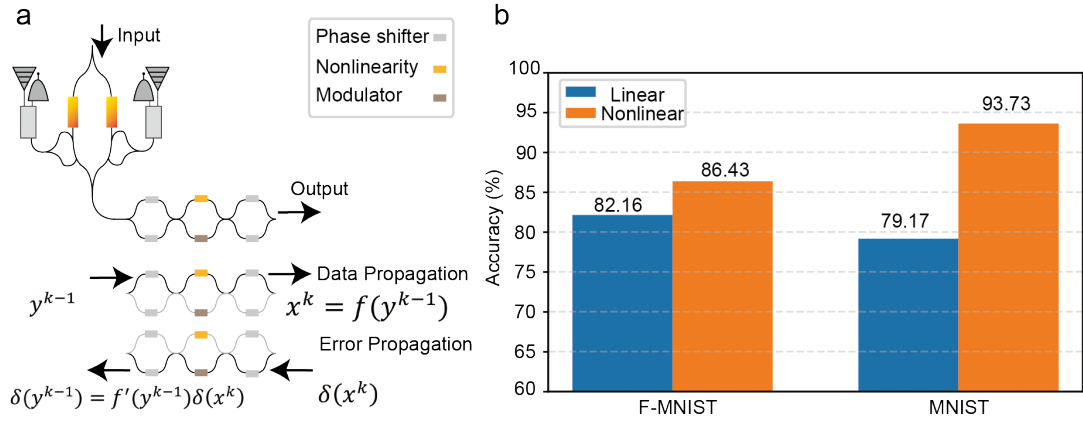
Supplementary Figure 4. Experimental calibration system. **a**, To align the phase profiles of the STR units on-chip, an on-axis interference system was constructed. The input beam from the laser source was split into two paths: one directed into the INSPIRE chip (device under test, DUT) and the other collimated into free space. The light emitted from the INSPIRE chip was then combined with the free-space plane wave, and the resulting interference pattern was captured by the camera. **b**, The phase was calculated by sweeping the on-chip phase and fitting the resulting intensity variation to a sinusoidal profile. The fitting curves of all 36 STR units are displayed, where the first 8 units in each column correspond to the STR units in a row. The final four units in the last row represent the results from the four STR units at the output positions.



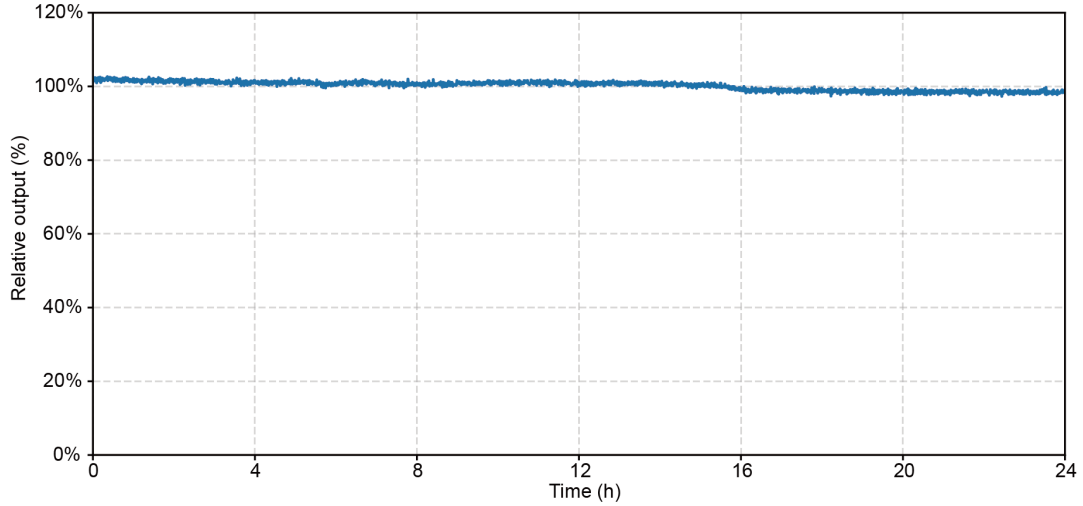
Supplementary Figure 5. Scattering layers for convolution operations. **a**, Schematic of the optical convolution backbone. **b**, Comparison of digital and optical convolution. The blue and red boxes represent digital and optical convolution results (normalized by a compensation factor 3.787), respectively. The error bar above the red box indicates the discrepancy between digital and optical results. **c**, Simulated intensity profile of the optical signal in a diffractive neural network. The input signal is encoded into the amplitude of the input waveguide mode, and the convolved results are obtained by recording the intensity at the output waveguide ports.



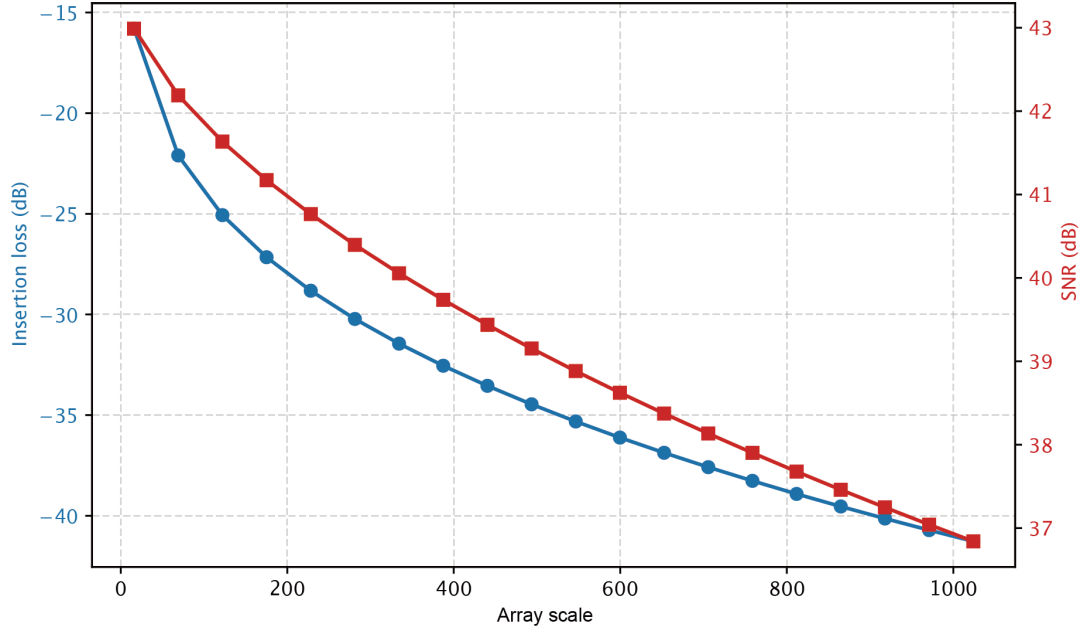
Supplementary Figure 6. Full trace of the designed 4×4 matrix in Fig. 2. Each row reflects the results of one entry in the matrix, showcasing the evolution of the training error, the output entry, and the gradients. The output gradients were scaled before being applied to update the STR units. Spikes observed in the training curve arise from random noise in the experimental system. To ensure convergence, a total of 500 iterations were performed.



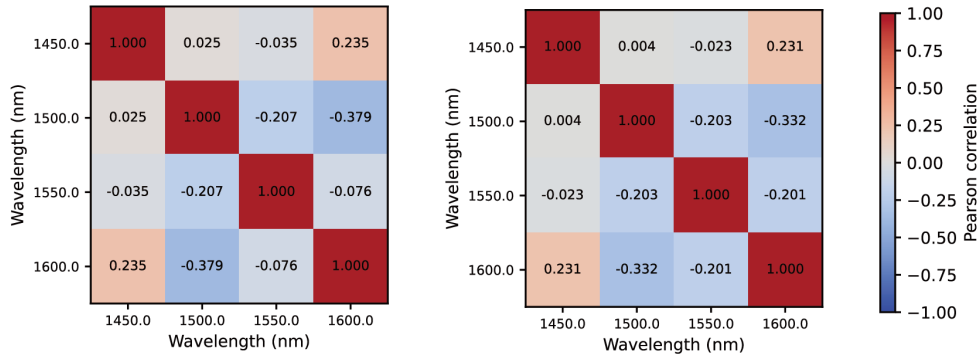
Supplementary Figure 7. Incorporation of nonlinearity in INSPIRE. **a**, The INSPIRE learning model with nonlinearity. **b**, The results on F-MNIST and MNIST datasets.



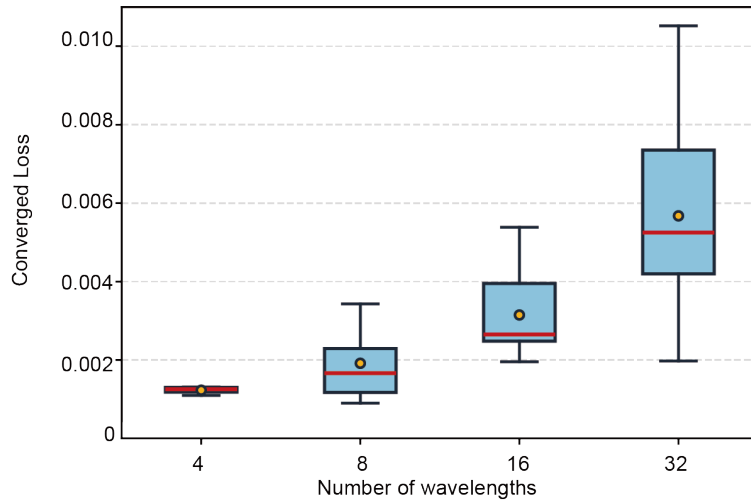
Supplementary Figure 8. The temporal stability of the measurements. We monitored the matrix output over 24 hours using a thermo-optical controller in the cleanroom temperature conditions. The relative error fluctuated between $\pm 2.683\%$. The fluctuation was mainly caused by ambient temperature variations from the air-conditioning system. With a more dedicated temperature controller, such as a constant-temperature chamber, the level of fluctuation could be further reduced. Even so, the current precision is sufficient for INSPIRE training.



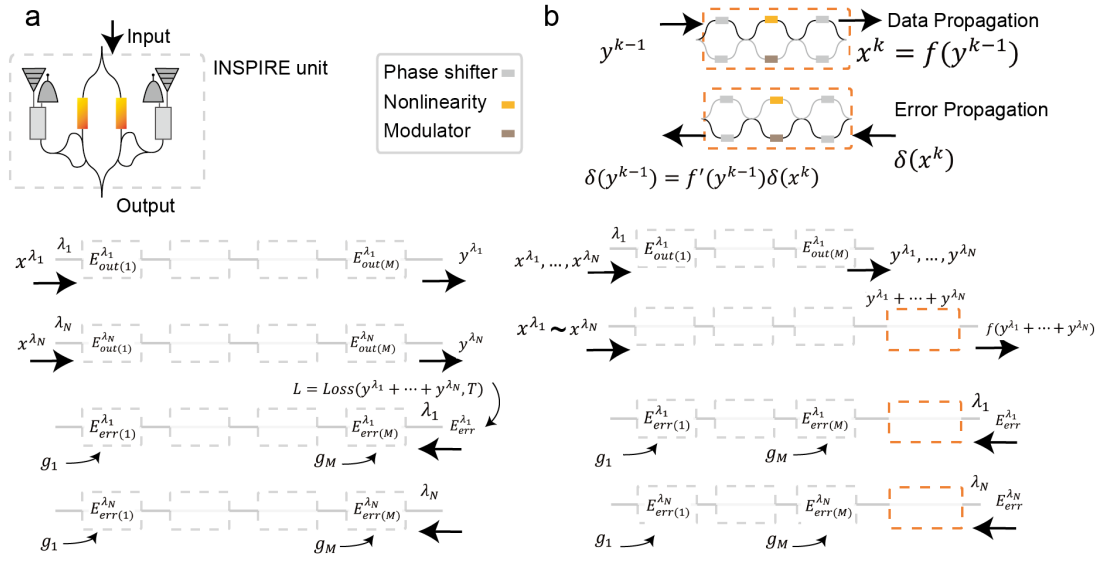
Supplementary Figure 9. The insertion loss and the signal-to-noise ratio (SNR) with respect to the array scale.



Supplementary Figure 10. The Pearson correlation matrix of the target and realized matrices in Fig. 3. We vectorized each wavelength-dependent 4×4 target matrix and the matrix realized by INSPIRE into a 16-dimensional vector and computed the Pearson correlation coefficient between all matrix pairs. The off-diagonal correlations are close to zero, ranging from approximately -0.38 to 0.24, indicating that the wavelength-dependent matrices are not trivially related by scaling or shared element-wise patterns.



Supplementary Figure 11. The converged loss of matrix fitting for varying number of wavelengths. The red lines and the orange dots indicate the median and average losses, respectively. We varied the number of wavelengths and evaluated the converged loss in multi-wavelength matrix fitting. For the four-wavelength case, the average mean squared error is 0.12%. As the number of wavelengths increases to 8, 16, and 32, the fitting error gradually increases to 0.19%, 0.31%, and 0.56%, respectively. These results indicate that the fitting error increases as the number of wavelengths scales up. In the current implementation, the system can reliably support four wavelengths, with each wavelength corresponding to an independently fitted (4×4) matrix. The degradation induced by the increase in the number of wavelengths could be further mitigated by incorporating a larger number of tunable STRs.



Supplementary Figure 12. INSPIRE training in a broadcast-and-weight wavelength-division-multiplexed system. a, Training in a linear system. Data and error propagation are performed independently for each wavelength channel. The wavelength-resolved gradients are then averaged and applied to update each INSPIRE unit. **b**, Training with nonlinear activation. In addition to the wavelength-resolved operations used in the linear system, the inputs from all wavelength channels are injected simultaneously during data propagation to evaluate the nonlinear activation. During error propagation, the error field passes through a modulator programmed with the derivative of the nonlinear function.

Supplementary Table 1. Comparison with existing photonic training methods.

Methods	In situ/In silico	Analytical/Approximation Gradients	Constraints/Applicability
INSPIRE (This work)	In situ	Analytical gradients	Applied to scattering media and non-unitary circuits
Perturbation-based Method ²⁻⁴	In situ	Approximated gradients	MZI meshes and diffractive circuits
Hybrid Method ^{5, 6}	In situ (forward) & In silico (backward)	Approximated gradients	Planar waveguide
In-situ Backpropagation ⁷	In situ	Analytical gradients	Unitary circuits
Fully forward mode Training ⁸	In situ	Analytical gradients	Symmetric circuits

Supplementary References

1. Instruments, T. Transimpedance Amplifier Noise Considerations. *Technical Articles*, (2021).
2. Bandyopadhyay, S. et al. Single-chip photonic deep neural network with forward-only training. *Nature Photonics* **18**, 1335-1343 (2024).
3. Wu, B. et al. Scaling up for end-to-end on-chip photonic neural network inference. *Light: Science & Applications* **14**, 328 (2025).
4. Zhang, J. et al. Highly integrated all-optical nonlinear deep neural network for multi-thread processing. *Advanced photonics* **7**, 046003-046003 (2025).
5. Wu, T., Menarini, M., Gao, Z. & Feng, L. Lithography-free reconfigurable integrated photonic processor. *Nature Photonics* **17**, 710-716 (2023).
6. Wu, T., Li, Y., Ge, L. & Feng, L. Field-programmable photonic nonlinearity. *Nature Photonics* **19**, 725-732 (2025).
7. Pai, S. et al. Experimentally realized in situ backpropagation for deep learning in photonic neural networks. *Science* **380**, 398-404 (2023).
8. Xue, Z., Zhou, T., Xu, Z., Yu, S., Dai, Q. & Fang, L. Fully forward mode training for optical neural networks. *Nature* **632**, 280-286 (2024).