Research Article

Compact planar-waveguide integrated diffractive optical neural network chip

Jianan Feng,[a,b] Chang Li,[a,b] Dahai Yang[c,d], Yang Liu,[a,b] Jianyang Hu[a,b], Chen Chen,[e] Yiqun Wang,[e] Jie Lin,[a,f,*] Lei Wang,[b] and Peng Jin[a,b,*]

[a]Harbin Institute of Technology, Ministry of Education, Key Laboratory of Micro-systems and Micro-structures Manufacturing, Harbin, China
[b]Harbin Institute of Technology, School of Instrumentation Science and Engineering, Harbin, China
[c]Great Bay University, School of Physical Sciences, Dongguan, China
[d]Great Bay University, Great Bay Institute for Advanced Study, Dongguan, China
[e]Chinese Academy of Sciences, Suzhou Institute of Nano-Tech and Nano-Bionics, Suzhou, China
[f]Harbin Institute of Technology, School of Physics, Harbin, China

Abstract. Diffractive optical neural networks (DONNs) have exhibited the advantages of parallelization, high speed, and low consumption. However, the existing DONNs based on free-space diffractive optical elements are bulky and unsteady. In this study, we propose a planar-waveguide integrated diffractive neural network chip architecture. The three diffractive layers are engraved on the same side of a quartz wafer. The three-layer chip is designed with 32-mm^3 processing space and enables a computing speed of 3.1×10^9 Tera operations per second. The results show that the proposed chip achieves 73.4% experimental accuracy for the Modified National Institute of Standards and Technology database while showing the system's robustness in a cycle test. The consistency of experiments is 88.6%, and the arithmetic mean standard deviation of the results is ~4.7%. The proposed chip architecture can potentially revolutionize high-resolution optical processing tasks with high robustness.

Keywords: optical computing; diffractive neural network; planar waveguide; high robustness.

1 Introduction

Artificial neural networks (ANNs) are rapidly developing and are widely utilized in many fields, such as computer vision,[1] natural language processing,[2] medical diagnosis,[3] and decision-making.[4] Although ANNs have notably improved performance at the algorithmic level, these tasks are fundamentally limited by the energy consumption and computing speed of computers.[5]

Recently, optical neural networks (ONNs) have gained increasing attention owing to their low power consumption, low processing latency, and high computational bandwidth for solving the above problems.[6–14] Various ONN implementations have been proposed, including coherent photonic integrated circuits,[15–21] phase-change materials,[22–24] diffractive optical processors,[25–28] dielectric metasurfaces,[29–31] and optical delay lines.[32] Among these ONNs, diffractive optical neural networks (DONNs) have attracted increasing interest because of their large computational scales.[33] However, the existing DONNs based on discrete diffractive optical elements (DOEs) are bulky and unsteady.[34–46]

In this study, we propose a compact planar-waveguide integrated DONN chip. The three diffractive layers are engraved on the same side of a quartz wafer and enable high-precision alignment. Meanwhile, the optical field avoids noise in the transparent waveguide. A three-layer chip is designed with a 32-mm^3 processing space. The compact architecture enables a computing speed of 3.1×10^9 Tera operations per second (TOPS). The three-layer chip achieves 73.4% experimental accuracy for the Modified National Institute of Standards and Technology (MNIST) database while showing the system's robustness in a cycle test. The consistency of the experiments is 88.6%, and the arithmetic mean standard deviation of the classification results is 4.7%. Furthermore, the chip can be combined with a complementary

*Address all correspondence to Jie Lin, linjie@hit.edu.cn; Peng Jin, P.Jin@hit.edu.cn

metal-oxide-semiconductor to achieve higher integration. This work provides a novel high-density integration solution with high robustness for high-resolution optical processing tasks.

2 Methods

2.1 Oblique Forward Propagation Model

For the mainstream DONNs, the forward propagation model is based on the angular spectrum (AS) theory of diffraction and fast Fourier transform. According to the AS theory, the diffraction field of a beam propagating in free space through a distance z can be expressed as

$$U_{\text{out}}(x, y) = \mathcal{F}^{-1}\{\mathcal{F}\{U_{\text{in}}(x, y)\}H(\xi, \eta)\}, \tag{1}$$

where $\mathcal{F}\{\}$ and $\mathcal{F}^{-1}\{\}$ are Fourier transform and inverse Fourier transform, respectively. $U_{\text{in}}(x, y)$ and $U_{\text{out}}(x, y)$ are the complex amplitude distributions of the light field on the input and output planes, respectively. $H(\xi, \eta)$ is a transfer function, which can be expressed as

$$H(\xi, \eta) = \exp\left[jkz\sqrt{1 - \lambda^2(\xi^2 + \eta^2)}\right], \tag{2}$$

where $k = 2\pi/\lambda$ and λ is the wavelength in the air.

For the proposed chip architecture, the diffraction in waveguide space is considered. Reflections are introduced from the l'th diffractive layer to the $l + 1$'th diffractive layer to match the waveguide space. The l'th diffractive layer with a transmittance function of $t(x_i, y_i)$ is obliquely illuminated with the incident angle θ. i'th represents the neuron located at (x_i, y_i) of layer l. The center of the input plane is at the origin of the coordinate system, whereas the regions of interest on the output plane are not located in the origin. Consequently, it is necessary to set the observation window after the oblique transmission during the simulations. During one simulation, the diffracted field can be expressed as

$$U_{\text{out}}(x, y) = \mathcal{F}^{-1}\{\mathcal{F}\{t(x_i, y_i)\exp(jk_w x \sin \theta)\}H(\xi, \eta)\}, \tag{3}$$

where $k_w = 2\pi/\lambda_w$ and λ_w is the wavelength in the waveguide space. In Eq. (3), a default setting is that the coordinates (x, y) on the output plane are the same as the coordinates (x_i, y_i) corresponding to the input plane. Reflections during transmission will cause the phase change. According to previous research,[47–49] we assume that the phase changes π at the reflecting interface. Therefore, the propagation in the waveguide space can be constructed through multiple diffractions and reflections.

2.2 Chip and Target Fabrication

The input targets are fabricated using laser direct writing on a soda glass substrate. The glass substrate is first cleaned using acetone and isopropyl alcohol. Using electron beam evaporation, the clear substrate is coated with a layer of chromium (Cr) with a thickness of a few hundred nanometers. After spin-coating positive photoresist and a prebake process, the handwritten digit patterns are exposed using laser direct writing. The exposed resist is removed using a developer, and the uncovered Cr is removed with chrome mordant. Any remnant resist is cleaned using acetone and isopropyl alcohol. Finally, 50 amplitude-encoded targets with a total of 10 categories are fabricated. Each category randomly selects five test targets in the MNIST test data set.

The phase value of each neuron is limited to 0 to 2π. The neuron phase value $\Delta\varphi$ is converted into a relative height map Δh [$\Delta h = \lambda\Delta\varphi/(2\pi\Delta n)$], where Δn is the refractive index difference between the substrate and air. The chip layers are fabricated onto a quartz wafer. After spin-coating photoresist and exposure, the exposed resist is stripped using a developer. Then, a magnetic neutral loop discharge etching is applied, which is followed by an oxygen plasma sizing treatment. This process is repeated until the chip layer structures are achieved. A more detailed description of the fabrication process is provided in the Appendix.

3 Results

The schematic of existing free-space DONNs is shown in Fig. 1(a). Discrete DOEs are independently distributed in free space, rendering the entire system bulky and unsteady. Moreover, the beams may be susceptible to the free-space noise. For a comparison, the schematic of the proposed chip is shown in Fig. 1(b). The DOEs are fabricated on the same quartz wafer. A reflective coating is fabricated on the back of the transparent substrate. The beam containing the target information is transmitted in the transparent waveguide space through diffractions and reflections. The output beam is received by a charge-coupled device (CCD) at the detection plane, and its intensity distribution is obtained. To fabricate the chip, all diffractive layers are integrated on the same surface. It enables high-level alignment between cascaded layers. Therefore, the chip design can achieve a compact and stable optical processing architecture.

A phase-only three-layer chip is designed for the classification task. To match the fabrication capability, the chip structure parameters are as follows. Each diffractive layer is 2 mm × 2 mm, which contains 250,000 (500 × 500) diffractive neurons. Each neuron unit is 4 μm in size. The horizontal interval of cascade layers is 1 mm. The transparent substrate thickness is 2 mm. The incidence angle of air is 60 deg. Therefore, the waveguide space for the beam propagation is only 32 mm³. The chip is trained using 55,000 amplitude-encoded handwritten digits. After training, the designed chip classifier tests 10,000 amplitude-encoded handwritten digits. The simulation classification accuracy is 75.4%. Some simulation results are shown in Fig. 2.

Subsequently, a three-layer chip is fabricated. The phase values of the diffractive layers are discretized to simplify fabrication. We construct the experimental optical path. In the
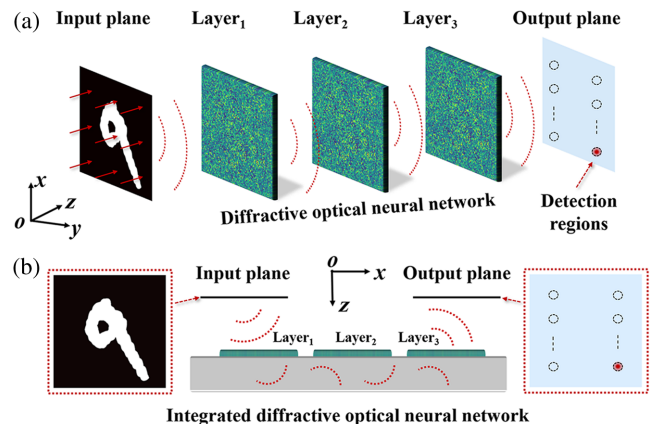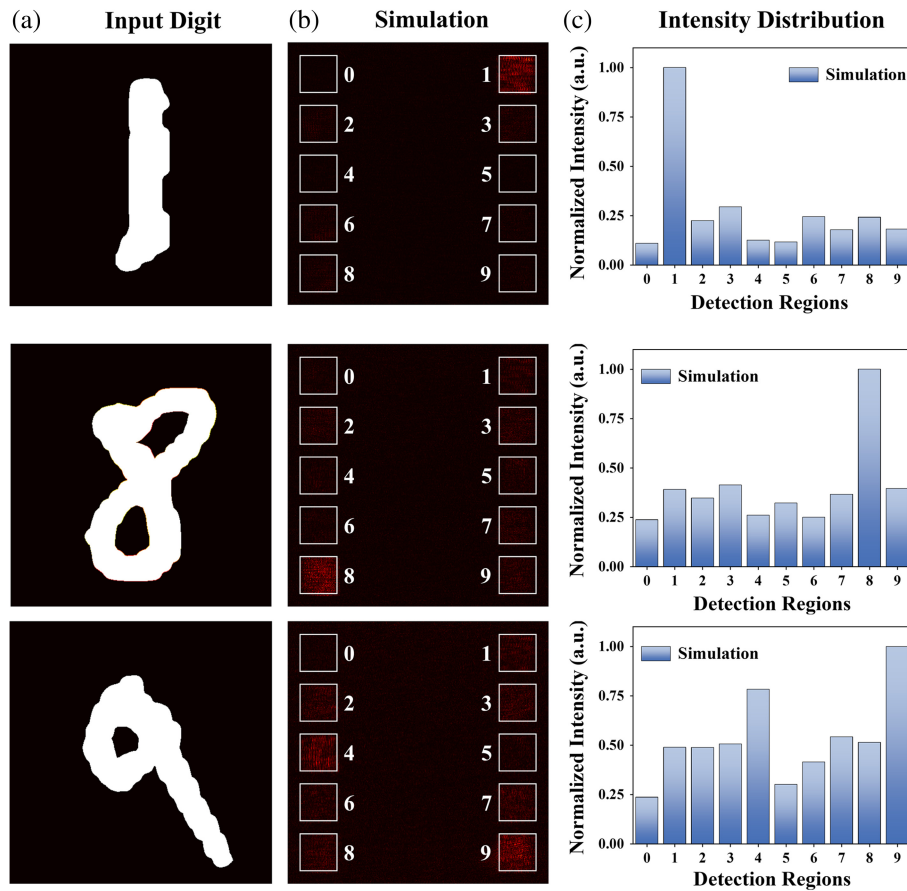

Fig. 1 Schemes of (a) existing DNNs and (b) the proposed chip.

Fig. 2 Simulation classification for the designed chip. (a) Input digits. (b) Simulation results. (c) Intensity distributions.

experiment, a He–Ne laser (25-STP-912-230, Melles Griot, Rochester, New York, United States) is collimated by lens1 and lens2. The wavelength of the He–Ne laser is 632.8 nm, and the power is 5 mW. A pinhole is used as a filter. The collimated beam illuminates the input plane. The intensity distributions in the output plane are detected by a CCD (DFK 33U×174, Sony, Minato, Tokyo, Japan). The fabricated device without reflective coating is shown in Fig. 3(c).

We randomly select 50 handwritten digits and fabricate them using laser direct writing. The fabricated handwritten digits are detailed in the Appendix. Some experimental results are shown in Fig. 4. The ability to classify different handwritten digits is assessed. Handwritten "1," "8," and "9" are chosen as the input targets, as shown in Fig. 4(a). The experimental output intensity distributions are shown in Fig. 4(b). The output intensity distributions are normalized considering the energy perturbation of the He–Ne laser. Then, we get the intensity ratios for 10 preset regions. As shown in Fig. 4(c), the maximal intensity appears at the preset region corresponding to the input handwritten digit label.

Furthermore, a 10-cycle test is performed to validate the reliability and stability of the three-layer chip. First, the chip is removed from the experimental optical path, whereas the rest of the experimental optical path remains unchanged. Subsequently, the chip is re-installed into the experimental optical path. The same test process is performed with the same test conditions to complete the 10-cycle test. For each handwritten digit, we calculate the arithmetic mean standard deviation of the intensity ratio. The error bars are shown in Fig. 5. For 500 test results, the arithmetic mean standard deviation of intensity ratio is 4.7%. The experimental classification accuracy is 73.4% in Fig. 6(b), and the experimental confusion matrix is shown in Fig. 6(a2). The statistical consistency of the 10-cycle test is 88.6% in Fig. 6(c). This is because there are some smaller errors (including rotational and deviation errors) between the input plane and the chip during the cycle test. The effect of alignment can be found in our previous work.[40]

Each layer of the designed three-layer chip contains 250,000 neurons. The cascaded layers are fully connected. The total number of operations is 1.25×10^{11}. The distance to complete one above session is ~12 mm. The time to complete one above session is $\sim 4 \times 10^{-11}$ s. Hence, the processing speed is $\sim 3.1 \times 10^{9}$ TOPS. In our previous free-space DONN work,[40] the distance to accomplish the same interlayer propagation was 10 cm. The propagation time of the proposed chip is about one-ninth of that of the free space diffractive neural network.

4 Conclusion

In this work, we proposed a compact planar-waveguide integrated diffractive neural network chip. Using micro-electromechanical system (MEMS) technology, the designed chip has realized a compact size of 32 mm³. Moreover, the compact architecture enables a computing speed of 3.1×10^{9} TOPS. The experimental accuracy is 73.4% in a 10-cycle test of 50 handwritten digits. The consistency of experiments is 88.6%. The
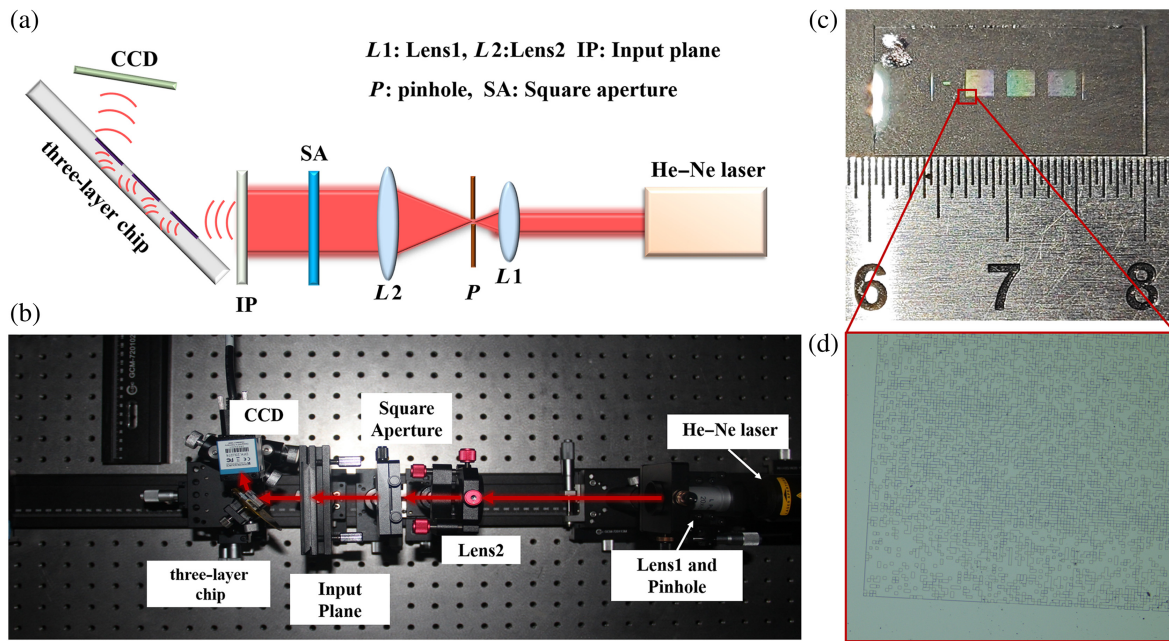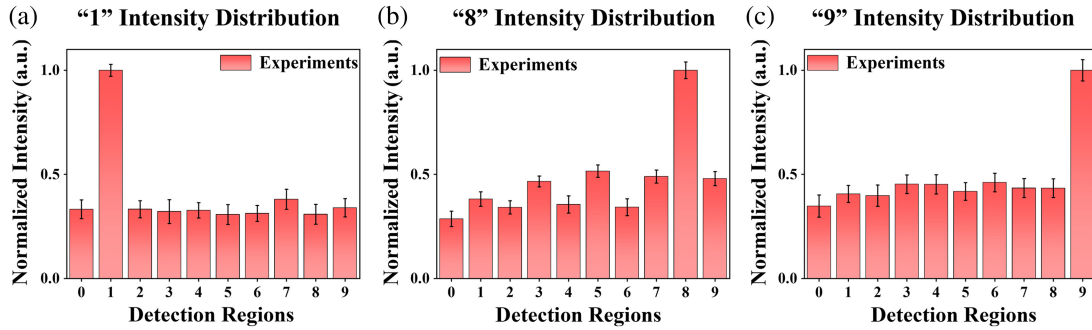
Fig. 3 Schemes of the experimental setup and fabricated chip. (a) Schematic diagram of the experimental setup. (b) Photo of the experimental setup. (c) The fabricated chip. (d) Partial enlarged view of the chip.
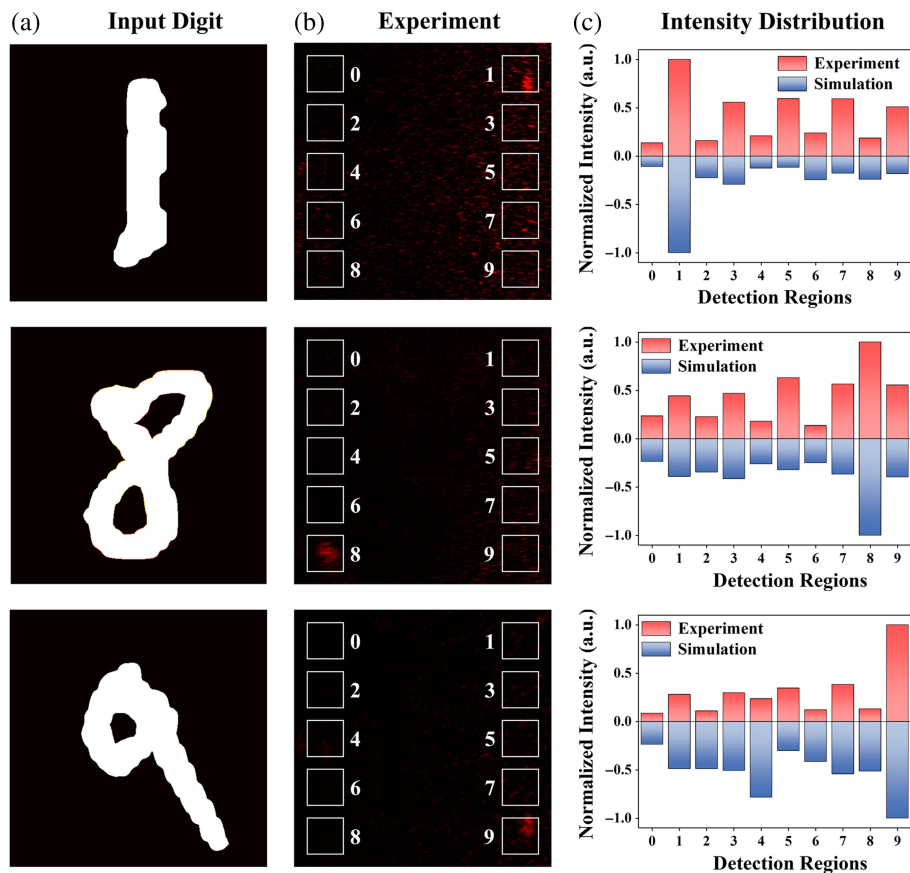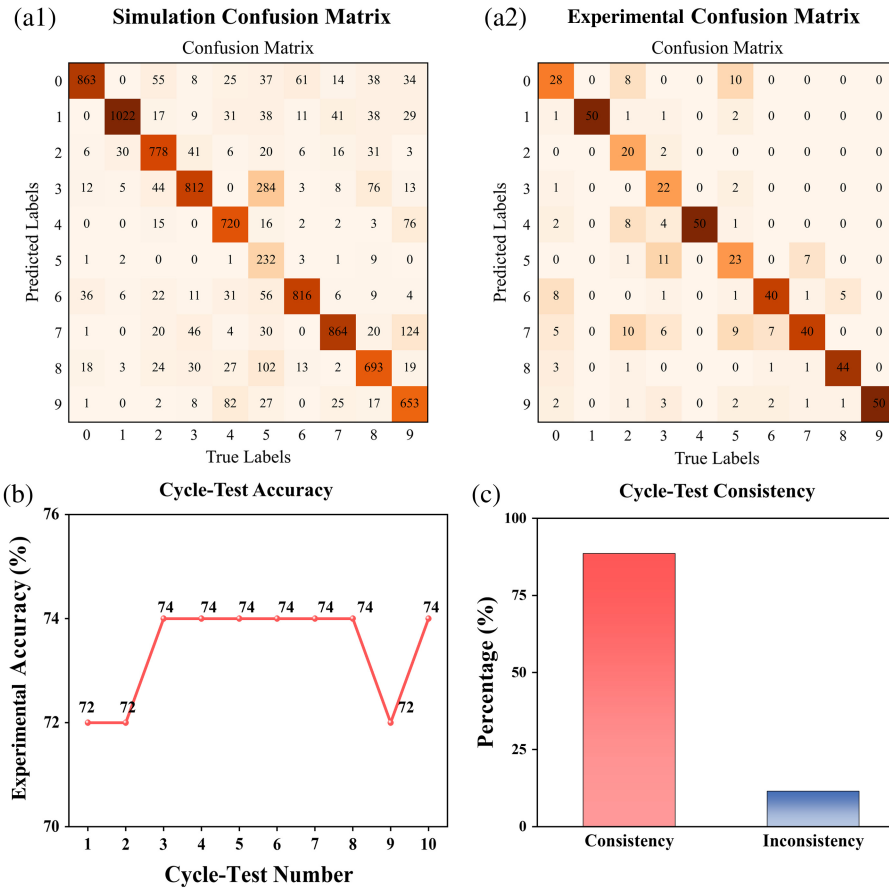

Fig. 4 Experimental classification for the designed chip. (a) Input digits. (b) Experimental results. (c) Intensity distributions.

Fig. 5 Cycle-test intensity results. (a) Intensity distribution of digit "1". (b) Intensity distribution of digit "8". (c) Intensity distribution of digit "9."

Fig. 6 Cycle-test consistency results. (a1) Simulation and (a2) experimental confusion matrices. (b) Accuracy of the 10-cycle test. (c) Consistency of the 10-cycle test.

arithmetic mean standard deviation is 4.7% for all 500 experimental normalized intensity distribution ratios. It will achieve an on-chip all-optical information processing unit with high alignment, high density, high reliability, and miniaturization, which provides a novel solution for high-resolution optical processing tasks with high robustness.

5 Appendix

5.1 Tensorflow-based Training

Here, a three-layer chip is constructed using an oblique forward propagation model. Between two adjacent layers, the propagation is completed by one reflection and two diffractions. At the detect plane, we obtain the output intensity distribution. Then, the mean square error is applied as the loss function. We aim to maximize the intensity of interest regions while minimizing the total intensity outside of all regions of interest. The trainable parameters in the chip are the modulation values for each layer, which are optimized using the backpropagation method of adaptive moment estimation (Adam) optimizer with a learning rate of 10^{-4}. To demonstrate the performance of the chip, we use 55,000 images in the MNIST data set for classification training. The designed chip is implemented using TensorFlow framework version 1.12.0 (Google Inc.) and Python (version 3.7.0). Using

a desktop computer (GeForce GTX 1660 graphical processing unit, AMD Ryzen 5 3600X CPU @ 3.8 GHz, 32 GB of random access Memory, and Microsoft Windows 10), a three-layer chip is trained for ∼20 h for 20 epochs of optimization iteration.

5.2 Fabrication of the Designed Three-layer Chip

In this paper, we train the designed three-layer chip to be between 0 and 2π. During the training, the phases are continuously distributed. To facilitate the actual processing, the trained phases are classified into four heights: 0, $\pi/2$, π, and $3\pi/2$. The designed three-layer chip is processed on a SiO_2 wafer. The etching depths for the SiO_2 corresponding to the phases 0, $\pi/2$, π, and $3\pi/2$ are 1038, 692, 346, and 0 nm, respectively. The fabrication steps are shown in Fig. 7.

(1) Step 1: cleaning
A 2 in (1 in = 2.54 cm) 500-μm-thick SiO_2 wafer that was polished on both sides was selected. The SiO_2 wafer was first ultrasonically cleaned with guaranteed reagent acetone (99.8%) and then transferred into guaranteed reagent isopropyl alcohol (99.8%). After the organic cleaning, the SiO_2 wafer was rinsed with deionized water for several minutes and finally dried with nitrogen gas. For further cleaning, we used an oxygen plasma surface treatment system to clean any organic impurities on the surface of the SiO_2 wafer.

(2) Step 2: spreading photoresist
Before spin-coating the photoresist, we pretreated the SiO_2 wafer with hexamethyl disilazane for 11 min. Based on the use of a positive photoresist, we chose a spin speed of 4000 revolutions per minute (rpm) for 30 s and soft bake at 95℃ for 90 s. Finally, the thickness of the photoresist was determined to be ∼900 nm, which is suitable for the following processing steps.

(3) Step 3: lithography and development
We carried out lithography and development for the first time in step 3. In step 3, we first used a photomask. We then used a Nikon Stepper i7 (Nikon, Tokyo, Japan) as our lithography machine with an exposure time of 600 ms and a focus of −1. Next, the exposed wafer was postbaked at 110℃ for 60 s to obtain a better development effect. Then, we developed the exposed wafer with a specific developer for 60 s. The developed wafer was then rinsed with deionized water for several seconds and finally dried with nitrogen gas. Up to this point, we have completed the first lithography and development step.

(4) Step 4: etching
We used a neutral loop discharge plasma etching system. For the plasma etching process, we chose octafluorocyclobutane (C_4F_8) and sulfur hexafluoride (SF_6) as the etching gas at an antenna RF power of 1200 W. For an etching time of around 36 s, we achieved an etching depth of 346 nm.

(5) Step 5: spreading photoresist
The etched SiO_2 wafer in step 4 was ultrasonically cleaned again. This is the second spin-coating process and is the same as in step 2.

(6) Step 6: lithography and development
This is the second lithography and development process and is the same as in step 3.

(7) Step 7: etching
This is the second etching process and is the same as in step 4.

(8) Step 8: spreading photoresist
The etched SiO_2 wafer in step 7 was ultrasonically cleaned again. This is the third spin-coating process and is the same as in step 2.

(9) Step 9: lithography and development
This is the third lithography and development process and is the same as in step 3.

(10) Step 10: etching
This is the third etching process and is the same as in step 4.

(11) Step 11: cleaning
Some positive photoresist remained on the surface, so the sample was cleaned again. This cleaning process was similar to step 1.
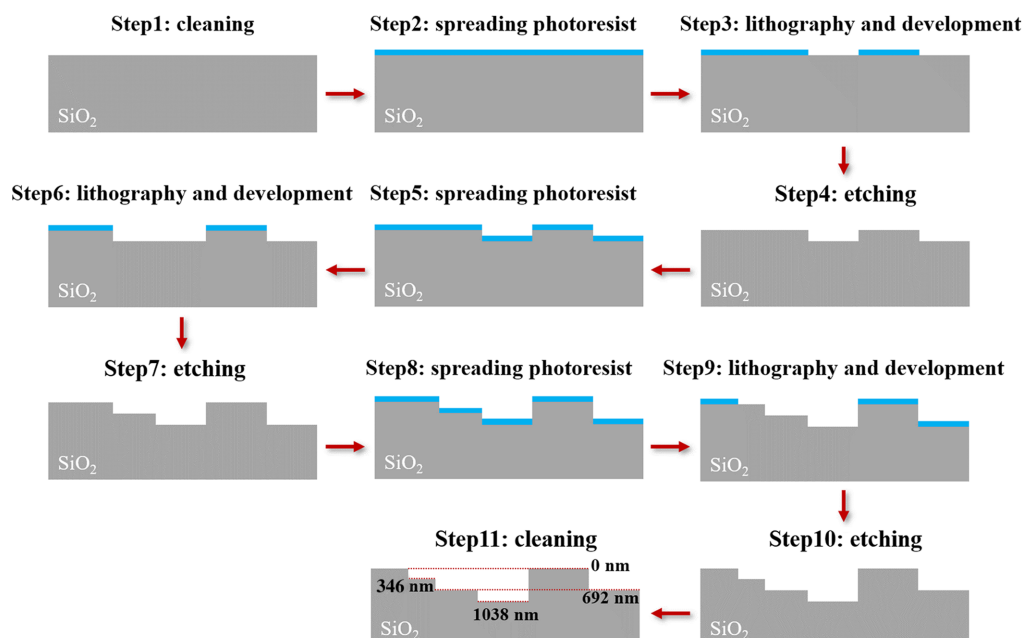

Fig. 7 Fabrication steps for the three-layer chip.

5.3 Simulation and Experimental Classification Results for 10 Categories of Handwritten Digits

Here, we show the classification results for 10 different categories of handwritten digits. As shown in Figs. 8(a) and 9(a),

handwritten digits "0 to 9" are used as input targets. The corresponding simulation and experimental classification results are shown in Figs. 8(b), 9(b), 8(c), and 9(c). In a similar way, we plot the error bars for the 10-cycle test in Figs. 8(d) and 9(d).
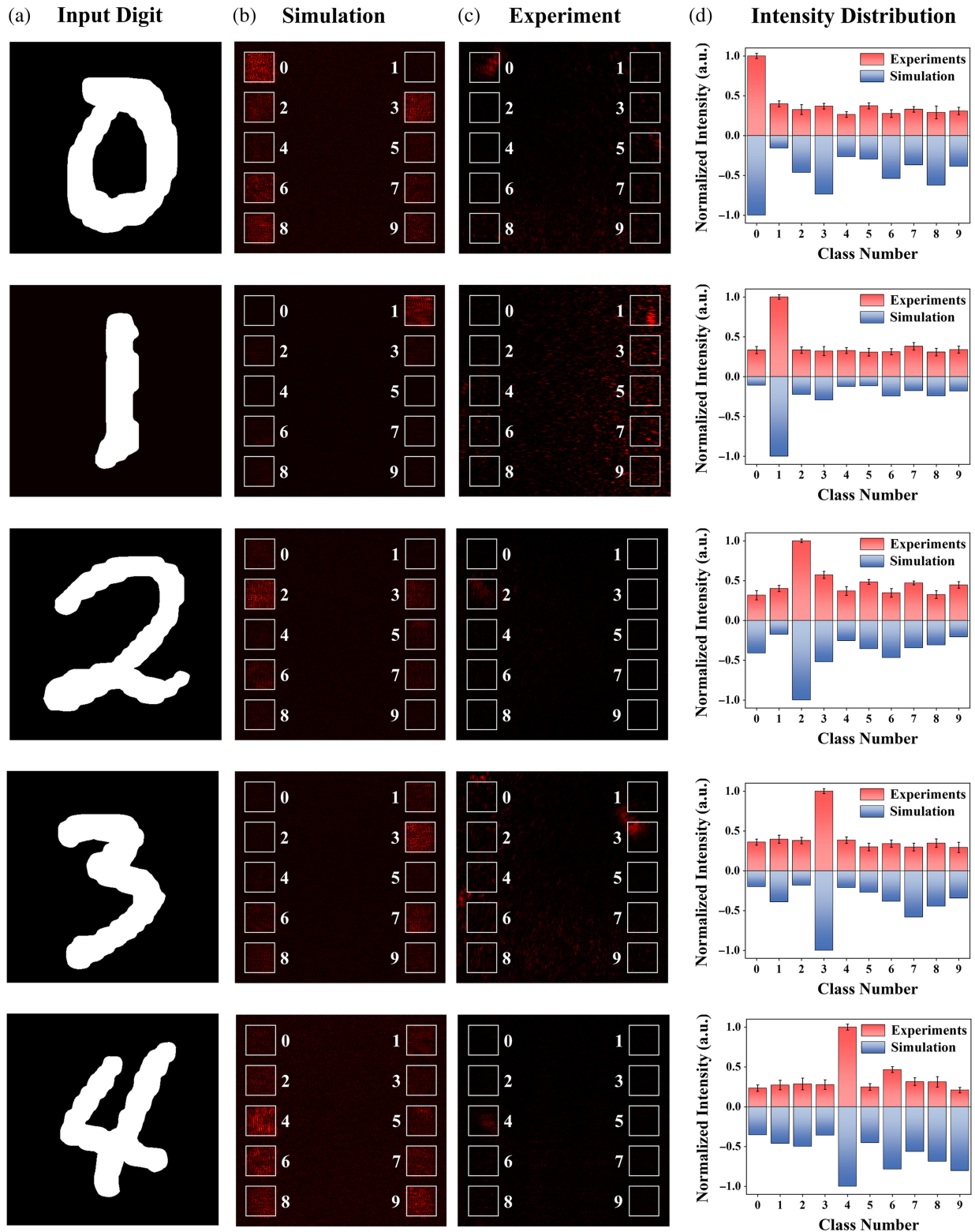
Fig. 8 Handwritten digit "0 to 4" classification for a three-layer chip. (a) Input digits. (b) Simulation results. (c) Experimental results. (d) Intensity distributions.
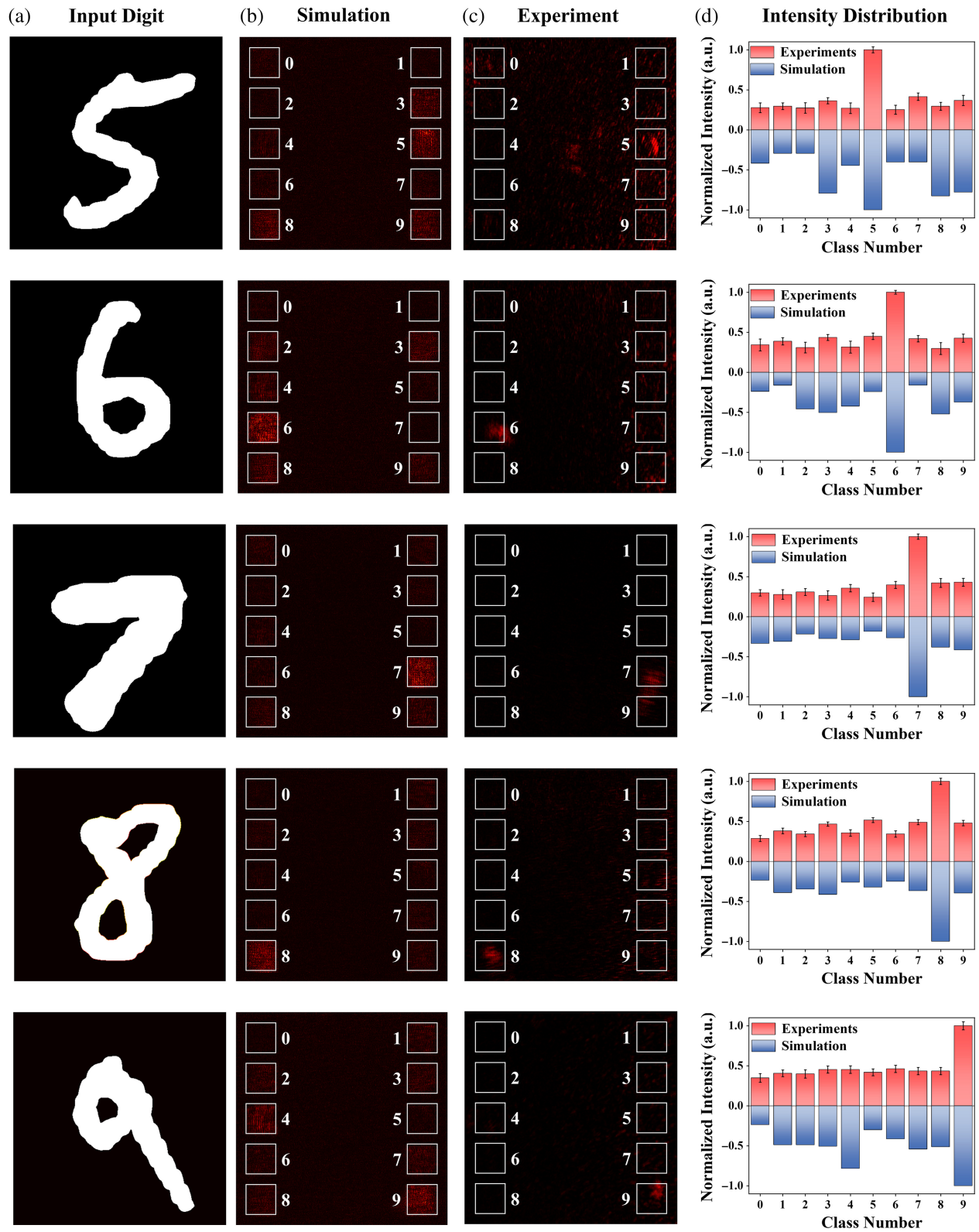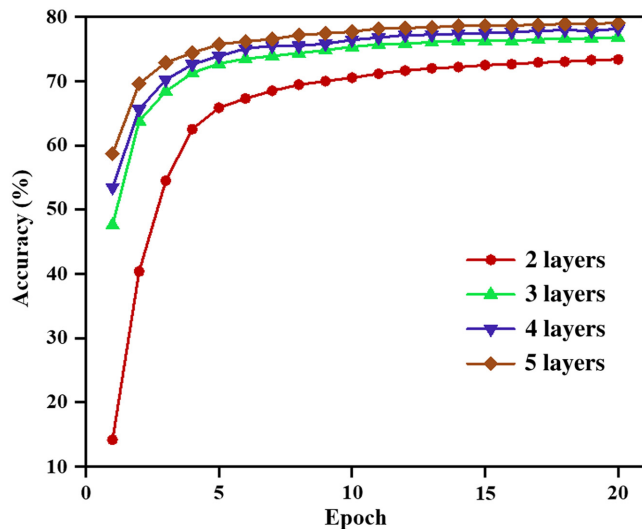
Fig. 9 Handwritten digit "5 to 9" classification for a three-layer chip. (a) Input digits. (b) Simulation results. (c) Experimental results. (d) Intensity distributions.

Fig. 10 Simulation classification accuracy for different numbers of layers for 10,000 test targets in the MNIST test data set.

5.4 Influence of the Layer Number on Recognition Accuracy

As shown in Fig. 10, the recognition accuracy of the chip is increasing for the increasing number of layers. Furthermore, the recognition accuracy was slightly changed from three to five layers. Therefore, a three-layer chip is analyzed.

5.5 Comparison between Different Architectures on TOPS

For optical diffractive neural networks, researchers have proposed many different architectures. We choose architectures that are similar to our work for better benchmarks. The comparison between different works and our work is shown in Table 1.

On the one hand, our work can achieve a more compact, integrated architecture compared with separate architectures. The computing power can be the same order of magnitude and can be further increased by expanding the number of neurons of phase layers.

On the other hand, our work has obviously higher computing power compared with the one-dimensional (1D) integrated architecture, and it can directly process the two-dimensional (2D) input objects. Moreover, our work has scalability while retaining stability.

5.6 Total Power Consumption of the Proposed Design

The experimental setup contains a He–Ne laser, a lens, a pinhole, a square aperture, an input plane, a three-layer chip, and a CCD. The He–Ne laser (light source) and CCD (detector) are active devices, and the others are passive devices.

The power of the He–Ne laser is 5 mW. The working current of the CCD is ~720 mA at 5 V, and the power of the CCD is ~3.75 W. The total power consumption of our scheme is ~3.755 W.

5.7 Heights of the Steps of the Fabricated Devices

The corresponding step heights of the diffractive layer are 346 nm. We measured the heights of the fabricated device steps using the confocal laser scanning microscope. The result of a measurement is shown in Fig. 11.

Three-dimensional (3D) microscope characterization of the step thickness of the proposed diffractive neural networks is shown in Fig. 11. The measured step heights are 302, 341, and 320 nm. The measurement error for the multistep photolithography-etching process is <30 nm. Although this kind of measurement error cannot be avoided, it does show a minor influence on the performance of diffractive networks.

5.8 Experimental Fabricated Targets

The fabricated targets are shown in Fig. 12. All 50 amplitude-encoded targets are randomly selected from the MNIST test dataset.

Disclosures

The authors declare no relevant financial interests.

Code and Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Table 1 Comparison between different architectures.

	Dimension	Size of neuron (μm)	Number of neurons in one layer	Propagation distance	Accuracy (%)	TOPS
Ref. 33	2D separation	400	200×200	12 cm	91.75	1.6×10^7
Ref. 40	2D separation	4	1000×1000	20 cm	84	6×10^9
Ref. 27	1D integration	2	186	500 μm	86.7	1.38×10^4
Our work	2D integration	4	500×500	1.2 cm	73.4	3.1×10^9
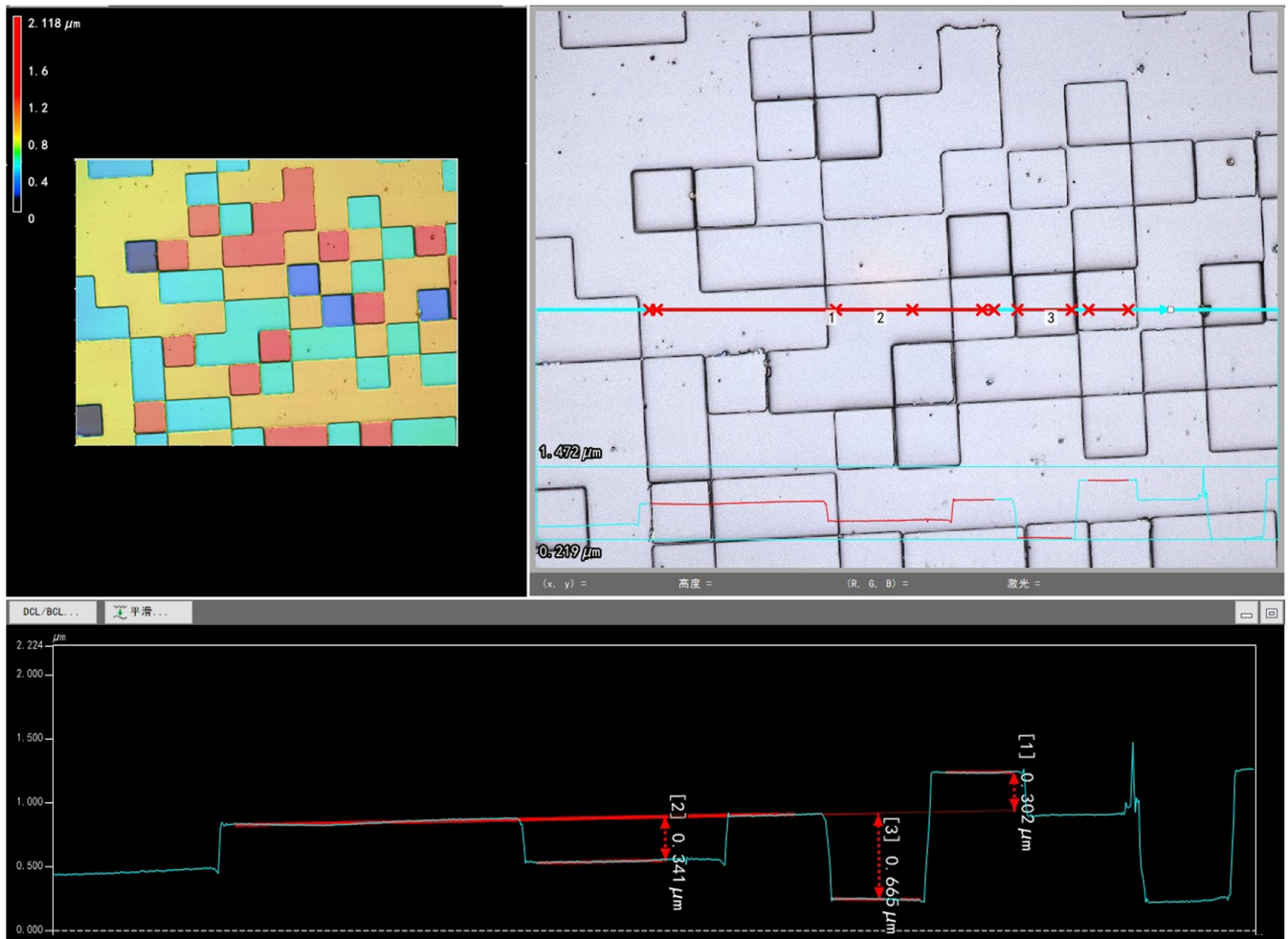
Fig. 11 3D microscope characterization of the step thickness of the proposed diffractive neural networks.
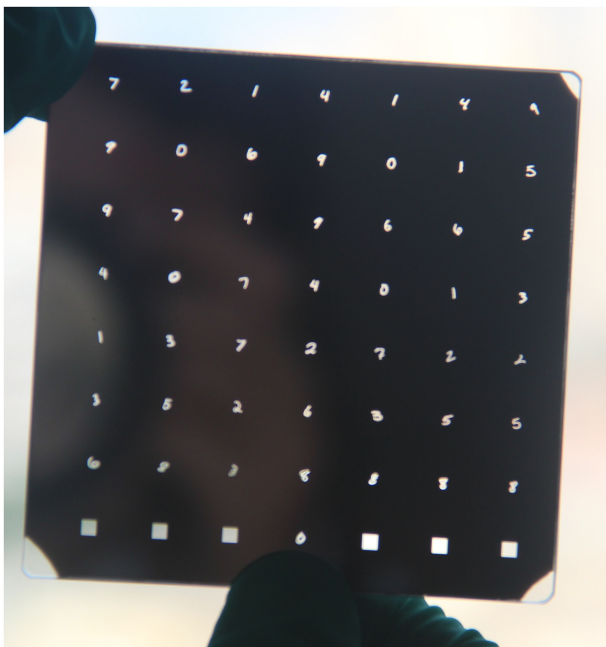
Fig. 12 Amplitude-encoded experimental fabricated targets.

References

1. A. Krizhevsky, I. Sutskever, and G. Hinto, "ImageNet classification with deep convolutional neural networks," *Commun. ACM* **60**, 84–90 (2017).
2. J. Hirschberg and C. Manning, "Advances in natural language processing," *Science* **349**, 261–266 (2015).
3. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).
4. I. Kruglov, O. Mishulina, and M. Bakirv, "Quantile based decision making rule of the neural networks committee for ill-posed approximation problems," *Neurocomputing* **96**, 74–82 (2012).
5. M. M. Waldrop, "The semiconductor industry will soon abandon its pursuit of Moore's Law. Now things could get a lot more interesting," *Nature* **530**, 144–147 (2016).
6. Q. M. Zhang et al., "Artificial neural networks enabled by nanophotonics," *Light Sci. Appl.* **8**, 42 (2019).
7. G. Wetzstein et al., "Inference in artificial intelligence with deep optics and photonics," *Nature* **588**, 39–47 (2021).
8. B. J. Shastri et al., "Photonics for artificial intelligence and neuromorphic computing," *Nat. Photonics* **15**, 102–114 (2021).
9. D. Perez et al., "Multipurpose silicon photonics signal processor core," *Nat. Commun.* **8**, 636 (2017).
10. L. Mennel et al., "Ultrafast machine vision with 2D material neural network image sensors," *Nature* **579**, 62–66 (2020).
11. J. M. Wu et al., "Analog optical computing for artificial intelligence," *Engineering* **10**, 133–145 (2022).

12. J. Spall, X. X. Guo, and A. I. Lvovsky, "Hybrid training of optical neural networks," *Optica* **9**, 803–811 (2022).
13. T. Yan et al., "All-optical graph representation learning using integrated diffractive photonics computing units," *Sci. Adv.* **8**, eabn7630 (2022).
14. B. Muminov and L. T. Vuong, "Fourier optical preprocessing in lieu of deep learning," *Optica* **7**, 1079–1088 (2020).
15. Y. C. Shen et al., "Deep learning with coherent nanophotonic circuits," *Natu. Photonics* **11**, 441–446 (2017).
16. T. W. Hughes et al., "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica* **5**, 864–871 (2018).
17. J. Feldmann et al., "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature* **569**, 208–214 (2019).
18. E. Khoram et al., "Nanophotonic media for artificial neural inference," *Photonics Res.* **7**, 823–827 (2019).
19. Z. Wang et al., "Integrated photonic metasystem for image classifications at telecommunication wavelength," *Nat. Commun.* **13**, 2131 (2022).
20. F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," *Nature* **606**, 501–506 (2022).
21. X. Y. Meng et al., "Compact optical convolution processing unit based on multimode interference," *Nat. Commun.* **14**, 3000 (2023).
22. J. Feldmann et al., "Calculating with light using a chip-scale all-optical abacus," *Nat. Commun.* **8**, 1256, (2017).
23. C. Rios et al., "In-memory computing on a photonic platform," *Sci. Adv.* **5**, eaau5759 (2019).
24. J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature* **589**, 52–58 (2021).
25. H. Zhang et al., "An optical neural chip for implementing complex-valued neural network," *Nat. Commun.* **12**, 457 (2021).
26. H. H. Zhu et al., "Space-efficient optical computing with an integrated chip diffractive neural network," *Nat. Commun.* **13**, 1044 (2022).
27. T. Z. Fu et al., "Photonic machine learning with on-chip diffractive optics," *Nat. Commun.* **14**, 70 (2023).
28. Y. H. Tang et al., "Device-system end-to-end design of photonic neuromorphic processor using reinforcement learning," *Laser Photonics Rev.* **17**, 2200381 (2023).
29. C. Qian et al., "Performing optical logic operations by a diffractive neural network," *Light Sci. Appl.* **9**, 59 (2020).
30. J. K. Weng et al., "Meta-neural-network for real-time and passive deep-learning-based object recognition," *Nat. Commun.* **11**, 6309 (2021).
31. C. Liu et al., "A programmable diffractive deep neural network based on a digital-coding metasurfaces array," *Nat. Electron.* **5**, 113–122 (2022).
32. X. Y. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural network," *Nature* **589**, 44–51 (2021).
33. X. Lin et al., "All-optical machine learning using diffractive deep neural network," *Science* **361**, 1004–1008 (2018).
34. S. M. Jiao et al., "Optical machine learning with incoherent light and a single-pixel detector," *Opt. Lett.* **44**, 5186–5189 (2019).
35. T. Yan et al., "Fourier-space diffractive deep neural network," *Phys. Rev. Lett.* **123**, 023901 (2019).
36. Y. Luo et al., "Design of task-specific optical systems using broadband diffractive neural networks," *Light Sci. Appl.* **8**, 112 (2019).
37. Y. Zuo et al., "All-optical neural network with nonlinear activation functions," *Optica* **6**, 1132–1137 (2019).
38. H. K. Dou et al., "Residual D^2NN: training diffractive deep neural networks via learnable light shortcuts," *Opt. Lett.* **45**, 2688–2691 (2020).
39. T. K. Zhou et al., "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nat. Photonics* **15**, 367–373 (2021).
40. H. Chen et al., "Diffractive deep neural networks at visible wavelengths," *Engineering* **7**, 1483–1491 (2021).
41. S. S. Rahman et al., "Ensemble learning of diffractive optical networks," *Light Sci. Appl.* **10**, 14 (2021).
42. E. Goi et al., "Nanoprinted high-neuron-density optical linear perceptrons performing near-infrared inference on a CMOS chip," *Light Sci. Appl.* **10**, 40 (2021).
43. M. Veli et al., "Terahertz pulse shaping using diffractive surfaces," *Nat. Commun.* **12**, 37 (2021).
44. A. Ryou et al., "Free-space optical neural network based on thermal atomic nonlinearity," *Photonics Res.* **9**, B128–B134 (2021).
45. J. X. Li et al., "Spectrally encoded single-pixel machine vision using diffractive networks," *Sci. Adv.* **7**, eabd7690 (2021).
46. J. S. Shi et al., "A physics-informed deep learning liquid crystal camera with data-driven diffractive guidance," *Commun. Eng.* **3**, 46 (2024).
47. Q. Jia et al., "Compensating the distorted OAM beams with near zero time delay," *Appl. Phys. Lett.* **121**, 011104 (2022).
48. Q. Jia et al., "Universal translation operator for Laguerre-Gaussian mode sorting," *Appl. Phys. Lett.* **121**, 191104 (2022).
49. Q. Jia et al., "Vector vortex beams sorting of 120 modes in visible spectrum," *Nanophotonics* **12**, 3955–3962 (2023).

Jianan Feng received his bachelor's degree in optoelectronic information science and engineering from the Harbin Institute of Technology (Weihai) in 2018. He has been a PhD student in instrument science and technology at the Harbin Institute of Technology since 2019. His research interests include optical diffraction neural networks.

Chang Li received his BS and MS degrees from the Harbin Institute of Technology in 2019 and 2021, respectively. He has been a PhD student in instrument science and technology at the Harbin Institute of Technology since 2021. His research interests include optical metasurfaces and MEMS devices.

Dahai Yang received his BS degree from Lingnan Normal University, his MS degree from the Harbin Institute of Technology (Weihai), and his PhD from the Harbin Institute of Technology in 2015, 2019, and 2024, respectively. He is currently a postdoc at the School of Physical Sciences of Great Bay University. His research interests include optical metasurfaces and OAM.

Yang Liu received his BS degree from the Harbin Institute of Technology (Weihai) in 2018 and his MS degree from the Harbin Institute of Technology in 2020. He has been pursuing a PhD in instrument science and technology at the Harbin Institute of Technology since 2020. His research interests include RF MEMS and millimeter-wave microsystem integration technology.

Jianyang Hu received his BS and MS degrees from the School of Physics and Optoelectronic Engineering at Harbin Engineering University in 2018 and 2021, respectively. He has been a PhD student in instrument science and technology at the Harbin Institute of Technology since 2021. His research interests include micro-nano optics and MEMS devices.

Chen Chen received his BEng and PhD degrees in instruments science and technology from the Harbin Institute of Technology in 2015 and 2022, respectively. He is currently a postdoc at the Nanofabrication Facility of Suzhou Institute of Nano-Tech and Nano-Bionics (Chinese Academy of Sciences). His research interests include optical metasurfaces and diffractive optics.

Yiqun Wang received his BEng and MEng degrees from the Changchun University of Science and Technology in 2005 and 2008, respectively, and his PhD in instruments science and technology from the Harbin Institute of Technology in 2021. He is currently a professor at the nanofabrication facility of Suzhou Institute of Nano-Tech and Nano-Bionics (Chinese Academy of Sciences). His research interests include micro- and nano-fabrication and integrated packages.

Jie Lin received his BS, MS, and PhD degrees in optics from the Harbin Institute of Technology in 2002, 2004, and 2007, respectively. He is currently a professor at the School of Physics, Harbin Institute of Technology. He has authored or coauthored more than 100 publications in international technical journals and conference proceedings. His research interests include the theory and application of micro-nano optics, metasurface, optical sensing, and imaging.

Lei Wang received his BEng, MEng, and PhD degrees in instrument science and technology from the Harbin Institute of Technology in 2000, 2002, and 2005, respectively. He is currently a professor at the School of Instrumentation and Engineering, Harbin Institute of Technology.

His current research interests include vibration isolation, ultraprecision motion control, measurement and instrumentation, and sensors and actuators.

Peng Jin received his BS degree in physics from Jilin University in 1994 and his MEng and PhD degrees in instrument science and technology from the Harbin Institute of Technology in 2001. He is currently a professor at the School of Instrumentation and Engineering, Harbin Institute of Technology. His research interests include fabrication and applications of micro-electro-mechanical systems, microwave passive components integration, and advanced processing technology for RF/microwave components.