

ARTICLE

Open Access

Anti-interference diffractive deep neural networks for multi-object recognition

Zhiqi Huang^{1,2}, Yufei Liu³, Nan Zhang^{1,2} , Zian Zhang^{1,2}, Qiming Liao^{1,2}, Cong He^{1,2}, Shendong Liu^{1,2}, Youhai Liu⁴, Hongtao Wang⁵ , Xingdu Qiao⁶, Joel K. W. Yang⁵ , Yan Zhang⁵ , Lingling Huang^{1,2}  and Yongtian Wang^{1,2} 

Abstract

Optical neural networks (ONNs) are emerging as a promising neuromorphic computing paradigm for object recognition, offering unprecedented advantages in light-speed computation, ultra-low power consumption, and inherent parallelism. However, most of ONNs are only capable of performing simple object classification tasks. These tasks are typically constrained to single-object scenarios, which limits their practical applications in multi-object recognition tasks. Here, we propose an anti-interference diffractive deep neural network (AI D²NN) that can accurately and robustly recognize targets in multi-object scenarios, including intra-class, inter-class, and dynamic interference. By employing different deep-learning-based training strategies for targets and interference, two transmissive diffractive layers form a physical network that maps the spatial information of targets all-optically into the power spectrum of the output light, while dispersing all interference as background noise. We demonstrate the effectiveness of this framework in classifying unknown handwritten digits under dynamic scenarios involving 40 categories of interference, achieving a simulated blind testing accuracy of 87.4% using terahertz waves. The presented framework can be physically scaled to operate at any electromagnetic wavelength by simply scaling the diffractive features in proportion to the wavelength range of interest. This work can greatly advance the practical application of ONNs in target recognition and pave the way for the development of real-time, high-throughput, low-power all-optical computing systems, which are expected to be applied to autonomous driving perception, precision medical diagnosis, and intelligent security monitoring.

Introduction

Deep learning techniques have proven to be effective in the classification and localization of objects in multiple scenarios¹. However, with the increasing complexity of application scenarios, current object recognition technologies face numerous challenges, such as simultaneously detecting multiple objects^{2–4}, especially when they occlude or overlap with each other. It is desired to develop a robust model for multi-object detection and instance segmentation capable of handling complex scenarios^{5–7}. Besides, certain applications require highly

precise recognition of fast-moving objects^{8–10}, which requires deep-learning-based high frame-rate processing capability and temporal information modeling ability^{11–14}. To meet the above requirements, tensor core processors for object recognition must deliver low latency, high throughput, and exceptional energy efficiency¹⁵. Traditional digital computers, however, face limitations in speed and energy due to Joule heating, electromagnetic crosstalk, and parasitic capacitance^{15–18}. Photonic technologies offer unparalleled advantages in the development of artificial intelligence hardware, providing solutions to overcome the bottlenecks of electronic systems¹⁹.

Optical neural networks (ONNs) represent a promising neuromorphic computing paradigm for object recognition, enabled by their unique advantages in light-speed computation, ultra-low power consumption, and inherent parallelism. As a key milestone in the advancement of photonic technology, ONNs are drawing increasing attention^{20–27}. In

Correspondence: Nan Zhang (nanzhang@bit.edu.cn) or Yan Zhang (yizhang@cnu.edu.cn) or Lingling Huang (huanglingling@bit.edu.cn)

¹Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China

²National Key Laboratory on Near-surface Detection, Beijing 100072, China

Full list of author information is available at the end of the article

These authors contributed equally: Zhiqi Huang, Yufei Liu, Nan Zhang

© The Author(s) 2026



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.



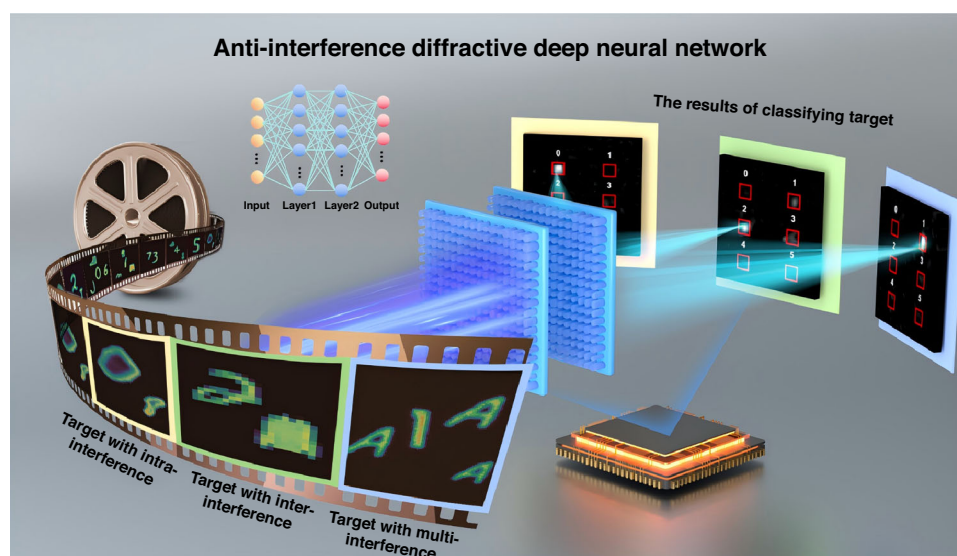


Fig. 1 Schematic illustration of Anti-Interference Diffractive Deep Neural Network (AI D²NN). The network can classify handwritten digits (0–5) in multi-object scenarios, including intra-class interference, inter-class interference, and dynamic interference

recent years, fundamental breakthroughs across diverse domains have accelerated the development of ONNs. Firstly, multi-dimensional multiplexing and interweaving techniques²⁸, including wavelength^{28–33}, polarization^{34–36}, and orbital angular momentum^{37,38}, enable massively parallel information processing³⁹. Meanwhile, energy-efficient materials and components, such as non-volatile phase-change material in a micro-ring resonator array^{40–43}, low-loss lithium niobate modulators^{44,45} and hybrid optoelectronic chips eliminating optical-electrical conversion⁴⁶, pave the way for the development of ONNs with significantly reduced system-level power consumption. Moreover, partially coherent optical neural networks can be implemented by reducing the coherence in either the temporal⁴⁷ or spatial domain⁴⁸, which has been shown to enhance experimental accuracy and robustness. Most notably, novel ONN architectures and training strategies are emerging^{49–52}, such as integrated diffractive-interference hybrid design⁵³, distributed computing architecture^{53,54}, and fully forward mode training methods⁵⁵, further enhancing inference performance and expanding functional diversity and complexity.

However, current ONNs are primarily designed for classifying a single target^{56,57}, which is far from the multi-object scenarios commonly encountered in real-world applications. Although some studies have investigated multiplexing schemes to facilitate multi-object classification³⁰, these schemes impose strict constraints on the location, size, and category of objects, thereby limiting the flexibility and practicality of the network. To overcome these limitations, several hybrid optoelectronic architectures have been proposed for dynamic scenes^{58,59}. However, these architectures require an electronic neural

network as a post-processing module to further process the low-dimensional features extracted by ONNs. The need for analog-to-digital conversion introduces latency and increases power consumption, thus compromising the intrinsic advantages of optical computing.

Here, an anti-interference diffractive deep neural network (AI D²NN) consisting of two transmissive diffractive layers is proposed, as schematically illustrated in Fig. 1. We employ distinct deep-learning-based training strategies to distinguish target with interference, and eliminate the impact of undesired interference on target's recognition result. In this work, targets are defined as handwritten digits 0–5 from Modified national institute of standards and technology (MNIST) dataset. To enhance the network's robustness against diverse forms of interference, an extensive interference dataset is utilized for training. This dataset includes intra-class interference from other handwritten digits 6–9 from MNIST dataset, inter-class interference derived from the Fashion-MNIST and EMNIST datasets. We also introduce dynamic interference by combining all of the aforementioned categories (40 categories in total) without any constraints on object location or size. The trained network can apply to multi-object scenarios, accurately and robustly recognizing targets and achieving a simulated blind testing accuracy of 87.4%. Furthermore, silicon-based metasurfaces are fabricated to physically implement the proposed AI D²NN. We establish a terahertz (THz) experimental platform with 0.85 THz as the incident light source. Using this setup, the AI D²NN achieves a blind testing accuracy of 86.7%, which is in good agreement with the numerical simulations.



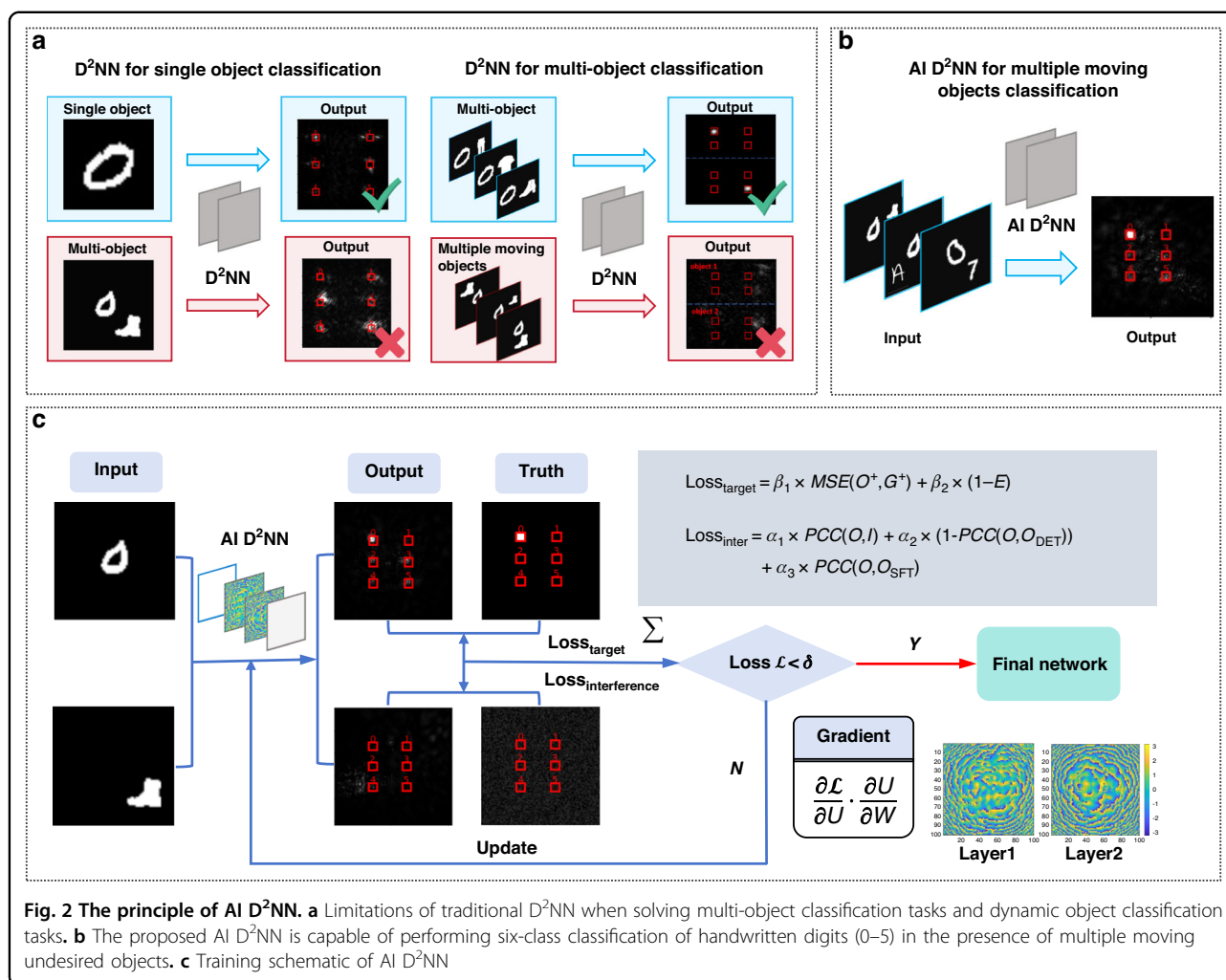


Fig. 2 The principle of AI D²NN. **a** Limitations of traditional D²NN when solving multi-object classification tasks and dynamic object classification tasks. **b** The proposed AI D²NN is capable of performing six-class classification of handwritten digits (0–5) in the presence of multiple moving undesired objects. **c** Training schematic of AI D²NN

The presented framework exhibits excellent scalability and can be scaled physically to operate across a broad wavelength range of electromagnetic waves, simply by scaling the diffractive features in proportion to the wavelength range of interest. In addition, it can be seamlessly integrated with optical multi-dimensional multiplexing technologies, enabling more flexible and higher-capacity parallel recognition tasks for multiple targets. Therefore, it can greatly advance the practical application of ONNs in target recognition and pave the way for the development of real-time, high-throughput, low-power all-optical computing systems, which are expected to be applied to autonomous driving perception, precision medical diagnosis, and intelligent security monitoring.

Results

Design of AI D²NN

Current ONNs demonstrate reliable performance for single-object recognition or spatially constrained multi-

object detection tasks, but exhibit substantial performance deterioration when applied to scenarios exceeding their functional scope, as shown in Fig. 2a. The proposed AI D²NN is capable of performing six-class classification of handwritten digits (0–5), even in the presence of undesired objects, as illustrated in Fig. 2b. To achieve this, the basic training schematic is shown in Fig. 2c. A customized training dataset was built by combining handwritten digits 0–5 with an interference set consisting of handwritten digits 6–9, fashions, and letters. Incorporating variability into the training dataset is intended to enhance the network’s robustness in recognizing target objects, enabling selective identification even in complex scenarios with interference across categories, sizes, and positions.

Next, an all-optical neural network was constructed, which is composed of an input layer, multiple diffractive layers and an output layer. The diffractive layers function similarly to hidden layers in fully connected electronic neural networks, processing input information in a linear

manner. Each unit within these layers performs complex amplitude modulation of the optical field, serving as a neuron node. Inspired by the Huygens-Fresnel diffraction principle, each neuron receives information in the form of a superposition of secondary optical fields generated by all neurons in the preceding layer, thereby tightly linking the layers and facilitating the hierarchical computation within the optical neural network. More detailed principal illustration of D²NN can be found in Supplementary Note 1.

During training, the data was first fed into the network, and the output was obtained via optical diffraction calculations. The loss value was computed based on the loss function, as guidance through back-propagation for fine-tuning the weights of each hidden layer. This procedure highlights the importance of loss function design, and for this task, the loss function was specifically tailored according to different labels.

For label 0–5, a conventional Mean Squared Error (MSE) loss function was employed to maximize the light intensity in the classification region aligned with the target category and minimize the intensity in other classification regions. Moreover, constraint on the light field intensity was imposed to improve the energy efficiency E within the correct classification region.

$$Loss_{target} = \beta_1 \times MSE(O^+, G^+) + \beta_2 \times (1 - E) \quad (1)$$

For label 6, the output optical field of interference is uniformly distributed as random noise outside the classification regions, which was achieved by designing a loss function based on the Pearson Correlation Coefficient (PCC). PCC quantifies the degree of linear dependence between two variables, with values ranging from -1 to 1. An absolute value approaching 1 indicates a strong linear correlation, while a value of 0 signifies no linear relationship, which can be expressed mathematically as:

$$PCC(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

The loss function for label 6 was designed as a weighted combination of three PCC metrics:

$$Loss_{inter} = a_1 \times PCC(O, I) + a_2 \times (1 - PCC(O, O_{DET})) + a_3 \times PCC(O, O_{SFT}) \quad (3)$$

Here, O and I mean optical field distribution in output plane and input plane separately. O_{DET} means optical field avoiding all classification regions. O_{SFT} means optical field

which shifts several pixels. $PCC(O, I)$ aims to minimize the similarity between the input and output, preventing the direct transmission of interference information to the output. $PCC(O, O_{DET})$ guides the spatial distribution of the output optical field to bypass the detection region when this value trends toward 1, and the $PCC(O, O_{SFT})$ is minimized to force the D²NN to generate uninterpretable noise-like output patterns.

Within each batch, the total loss was computed by first calculating the loss for each label separately and then summing these label-specific losses.

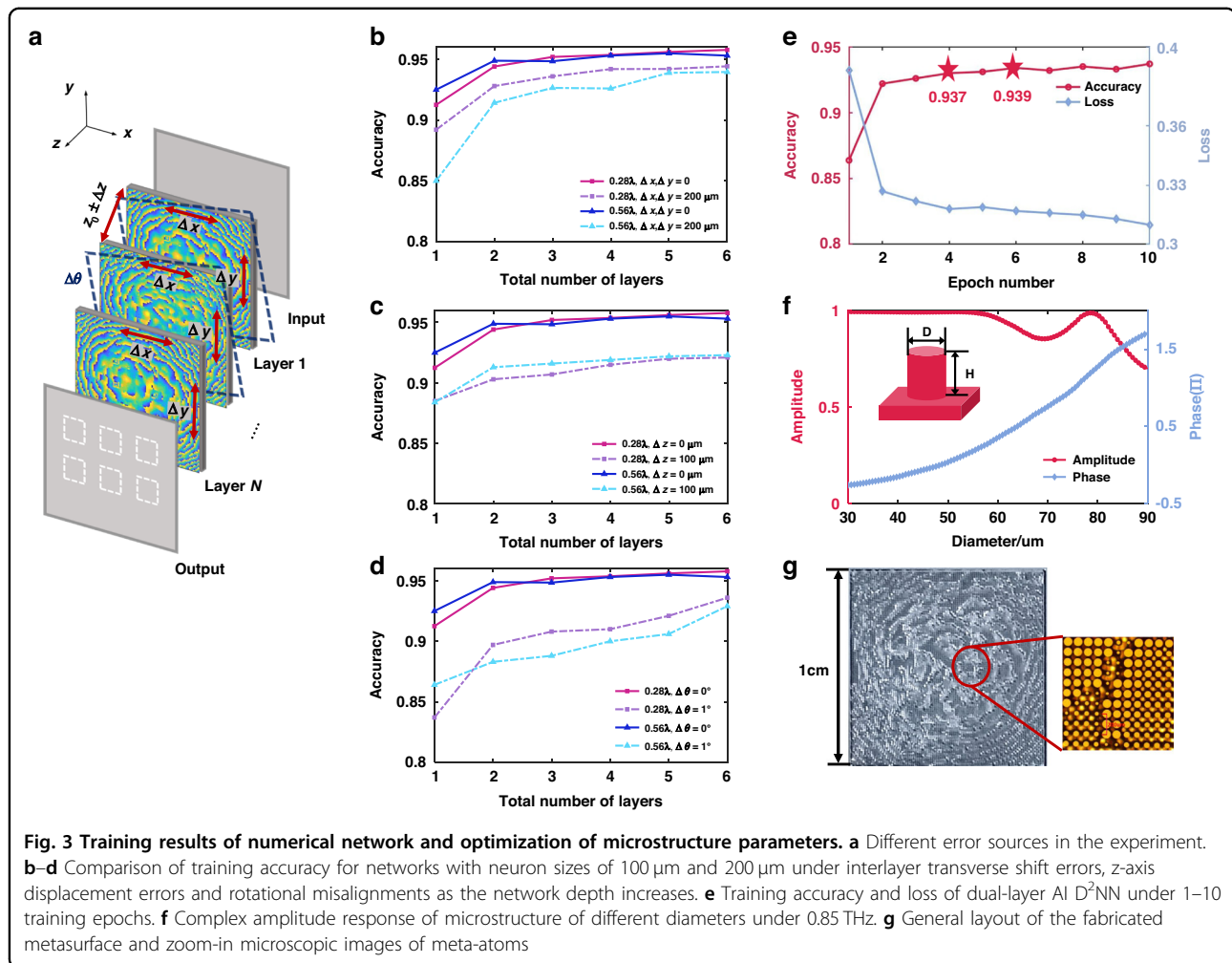
$$Loss = Loss_{target} + Loss_{inter} \quad (4)$$

The phase parameters of diffractive neurons were optimized using the backpropagation algorithm and stochastic gradient descent. After several epochs of training, the model converged and generated desired optical field distribution. More detailed information about training procedures and energy calculation methods are shown in Supplementary Note 2 and Supplementary Note 6.

To assess the classification accuracy of the network, a selective evaluation approach was designed. For digits 0–5, the classification was considered correct if the detection region with the maximum intensity corresponds to the correct label. For interference labeled 6, the recognition was deemed correct if both the mean and standard deviation of the optical output field were less than 0.2. This way provided an intuitive measure of the network's capability to suppress the influence of light field interference on digit classification results.

Additionally, other strategies were implemented to enhance the network's adaptability to experimental conditions. A 0.85 THz (352 μm) laser light source was employed for subsequent experiments. The impact of diffractive layer number, neuron sizes, and training iterations on network performance were systematically investigated. The diffractive surfaces, each comprising a 100×100 neuron array with respective neuron sizes of 100 μm and 200 μm , were compared under various experimental error conditions, including transverse shifts (within 200 μm), rotational misalignments (within 1 radian), and displacement errors along the z-axis (within 100 μm), as shown in Fig. 3a. When trained in the absence of random errors, both networks exhibit comparable classification accuracies. However, the performance of these two networks degrades to different extents when introducing aforementioned experimental errors during training, as illustrated in Fig. 3b–d. It can be concluded that the network with neuron sizes of 100 μm exhibits greater robustness compared to 200 μm counterpart. Considering fabrication constraints and alignment difficulties, a neuron size of 100 μm and a dual-layer configuration were selected for subsequent experiments. Subsequently, under the above network configuration, the





effect of training epochs (from 1 to 10) on classification accuracy was further assessed, as shown in Fig. 3e and Table S1 in Supplementary Note 2. After only 4 training epochs, the network achieves an accuracy exceeding 93.7%, which remains within the range of 93.7% to 94% in the subsequent epochs. Details about the combined effects of multiple experimental errors on network performance are available in Table S3 in Supplementary Note 3.

Design of the metasurfaces and elementary microstructures

As a proof of concept, the metasurface structures were designed for neuron phase modulation. The metasurface consists of high-impedance silicon cylinders deliberately fabricated on a silicon substrate. The electromagnetic response of each microstructure was calculated using the finite-difference time-domain (FDTD) method. The height of the microstructure was fixed at 240 μm , with a periodicity of 100 μm in both x and y directions. The incident wavelength was set at 0.85 THz.

To explore the phase modulation properties, the diameters of the microstructures were varied from 30 μm to 90 μm in 300 nm increments. This yielded the complex response coefficients of the microstructures under x-polarized incident light, as shown in Fig. 3f. The 16 types of structures with amplitude coefficients close to 1 and phase responses nearest to integer multiples of $\pi/8$ were selected. The specific structural parameters are listed in Table S4 in Supplementary Note 5. These structures were then assembled according to the neuron values of the trained diffractive neural network. Ultimately, two metasurfaces of 100 $\times$ 100 pixels were designed and fabricated for subsequent terahertz experiments. The overall distribution of the fabricated metasurface and the zoom-in optical microscopic images of metasurface are shown in Fig. 3g.

Anti-interference classification results

The proposed network is capable of accurately classifying the target in multi-object scenarios. Specifically, when the input contains both interference and target



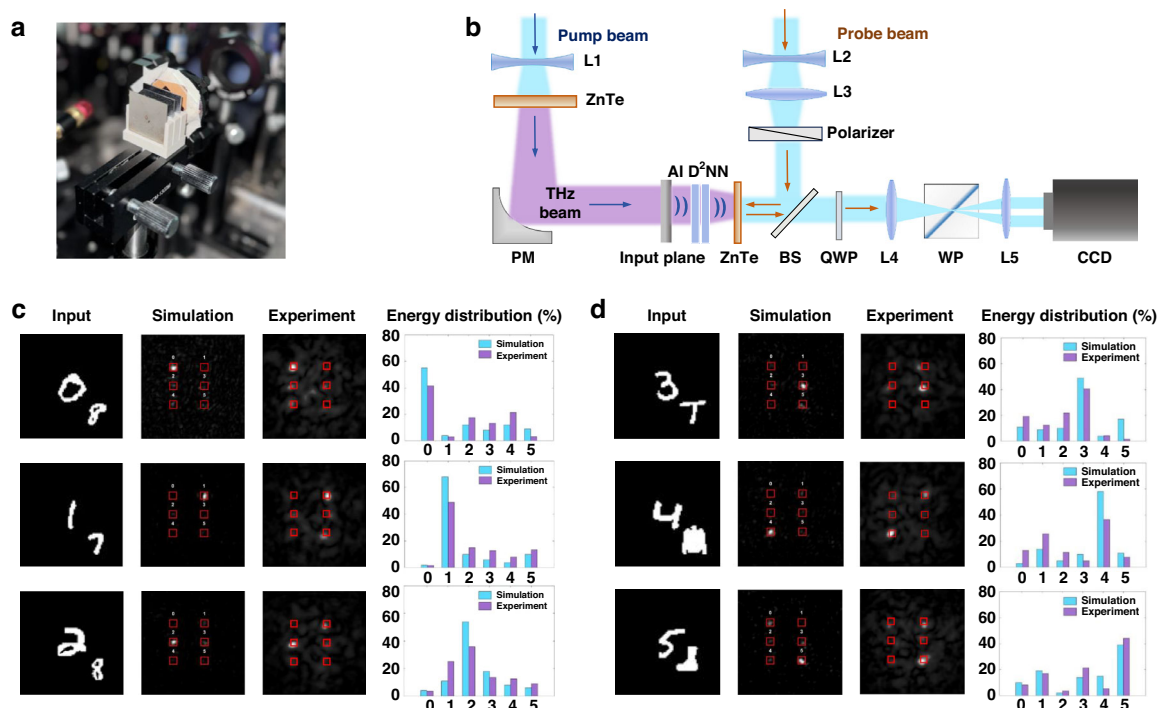


Fig. 4 Anti-Interference Classification Results. **a, b** Experimental ONN set up. **c** The results of AI D²NN classifying digits and intra-class interference. **d** The results of AI D²NN classifying digits and inter-class interference

objects, the AI D²NN identifies the target, as evidenced by the concentration of energy within a designated region on the output plane. To evaluate the network's performance, both numerical simulations and experimental evaluations were conducted.

In the numerical simulations, a test dataset of 6000 images containing digits and interference was built. These images were input into the trained network model to assess the classification accuracy of the primary digits. In the experiment, the input mask and two diffractive layers were assembled using 3D-printed holder, as shown in Fig. 4a, completing the physical cascading of the AI D²NN. A terahertz experimental platform was established, depicted in Fig. 4b, and 60 samples were tested to evaluate the accuracy of the physical network. Detailed experimental set-up and fabricated samples are presented in Fig. S1 in Supplementary Note 4.

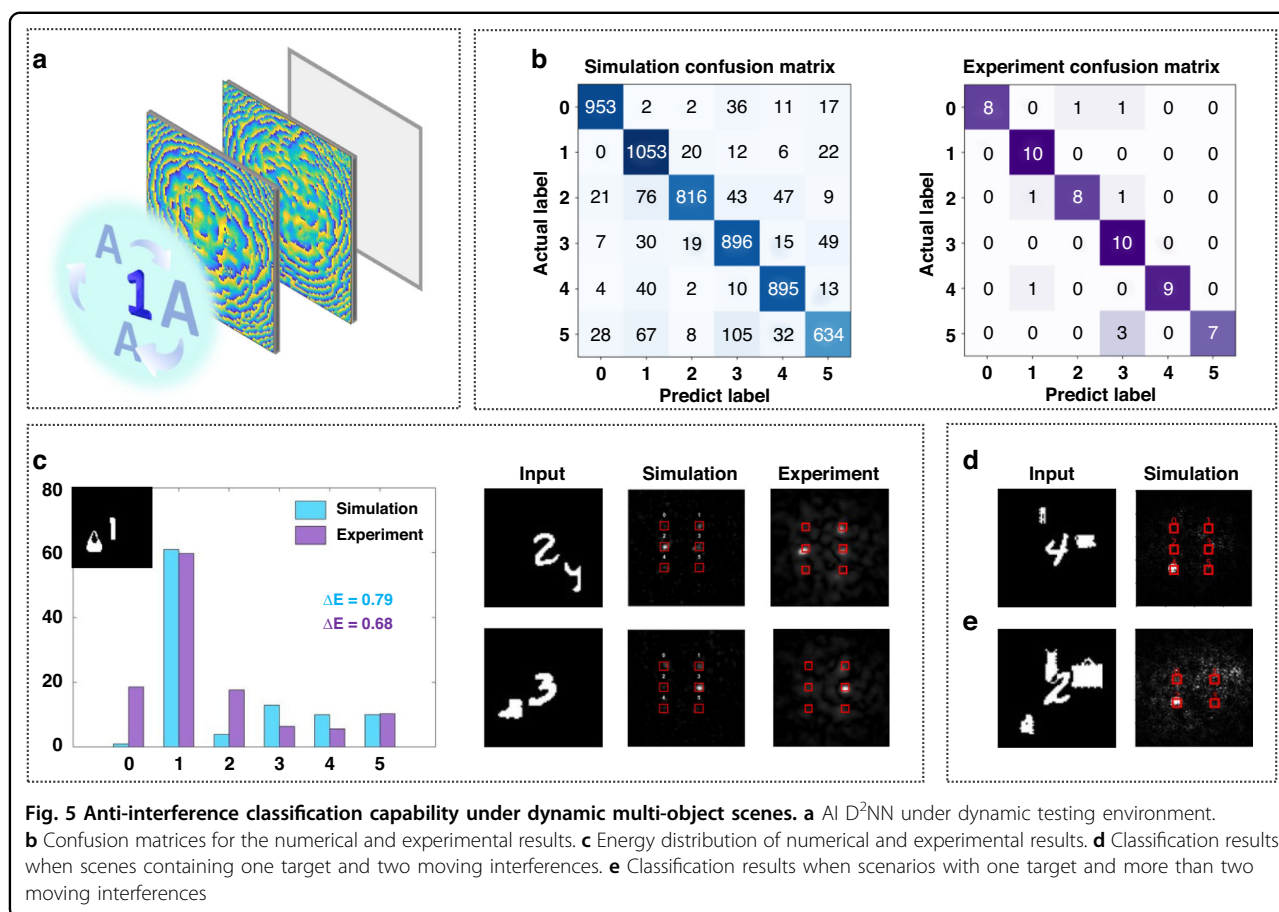
First, the network's classification capability in the presence of target digits 0–5 and intra-class interference digits 6–9 was evaluated. This case can mimic real-world industrial scenarios, such as automated inspection of specific types of components on a production line or the identification of vehicle types in transportation systems. The corresponding confusion matrix of the test dataset is shown in Fig. S3a in Supplementary Note 7, which indicates that the network achieved a recognition accuracy of 90.1% for digits 0–5 under scenes with intra-class

interference. Figure 4c demonstrates numerical simulation results and experimental results, which exhibit high consistency, proving that the delicately designed ONN metasurfaces possess the capability for target recognition.

Next, the network's classification capability was assessed when target and inter-class interference appear together. According to the confusion matrix in Fig. S3b in Supplementary Note 7, the network's recognition accuracy for digits 0–5 reaches 89.7% in the presence of inter-class interference. Both numerical simulation and experimental results are shown in Fig. 4d. The energy distribution demonstrates that the system effectively concentrates most of the energy at the correct classification region. Although light spots appear in other classification regions, the maximum optical field intensity in these regions is significantly lower than that in the correct region, with a ratio ranging from approximately 0.48 to 0.70. Furthermore, no significant light spots appear outside the classification regions, indicating that the network can accurately classify target while remaining resistant to the influence of other interference present in the scene.

To better simulate real-world scenarios, the dynamic multi-object scenes were constructed, where the target remains stationary while interfering objects allocate randomly in three-dimensional space. This configuration imposes natural challenges including objects clustering, partial occlusions, and scale variations, all arising





organically from the objects' movement. During the dataset construction process, we simulated the movement of interfering objects parallel to the detection plane by varying their positions in the image, and simulated their movement perpendicular to the detection plane by scaling the interfering objects. As shown in Fig. 5a, by randomly altering the sizes and positions of the interfering objects, a series of dynamic multi-object scenarios were generated. The test results for representative examples of correct and incorrect identifications are presented in Fig. 5b. In numerical simulations, the network achieved 87.4% accuracy on a test set of 6000 images. In experimental validation, it attained 86.7% accuracy with 10 samples per class, closely aligning with the simulation results and demonstrating the robustness of the design.

In addition, metrics such as signal-to-noise ratio (*SNR*) and the discrimination factor (ΔE) were incorporated into the network evaluation metrics. In this study, *SNR* is defined as the ratio of the maximum average intensity of the classification regions to the average intensity of the output plane outside the six detection regions, while ΔE represents the ratio of the difference between the highest and second-highest energies across the six detection regions to the highest energy. As presented in Fig. 5c,

numerical simulations yielded an *SNR* of approximately 31 and a ΔE value close to 0.8, while experimental measurements showed an *SNR* of around 23 and a ΔE of 0.68. Furthermore, Fig. 5d shows the network's output for the flexible test cases, demonstrating that the network can accurately classify the primary target regardless of the location and size of the interference. Besides, video streams were leveraged, which contains target (digit 4) and persistently moving interference as input. The corresponding output performance confirms the network's classification capability under dynamic scenes. The input and output videos are presented in Supplementary Video 1 and Fig. S5 in Supplementary Note 7.

The concept of transforming the optical fields of interfering objects into background noise through ONN is inherently extendable. Even as the number and categories of interference increase, the network still demonstrates excellent capability in classifying the primary target in multi-object scenarios. In theory, more complex scenarios for target recognition can be achieved by employing more diverse and extensive datasets, more rigorous loss functions and deeper network models during training phase. Therefore, we successfully constructed scenarios with one target and more than two interfering objects, in which the

positions and sizes of all interfering objects vary flexibly. The model's target recognition accuracy and the output optical field distribution were evaluated using numerical simulations, as shown in Fig. 5e, Figs. S6–S12 in Supplementary Note 7 and Figs. S15–S17 in Supplementary Note 8, which demonstrated AI D²NN has strong adaptability to complex scenarios. Lastly, a 50-frame video of the input and corresponding output is shown in Supplementary Video 2. The input scene contains a target (digit 2) and three interfering objects moving along randomly generated trajectories with varying sizes, further showcasing the network's scalability and generalization capability.

Discussion

Most existing ONNs are designed for single-object classification, and their performance deteriorates significantly when multiple objects present in a scene. To overcome this limitation, we proposed an anti-interference diffractive deep neural network for multi-object environments. Our approach enables the network to distinguish between target and interferences, and focus target's intensity onto designated detection region while transforming interferences' intensity into low-energy and uniformly distributed noise. This training strategy demonstrated strong robustness, achieving a numerical accuracy of 87.4% even in scenarios involving 6 types of targets and 40 types of dynamic interference. Furthermore, we constructed a terahertz-based diffractive neural network for experimental validation. The measured classification accuracy closely matched the simulation results, highlighting the network's potential for rapid and intelligent identification of primary targets in complex environments.

Future work can be developed regarding functional capability expansion, operational wavelength scaling and system-level integration. By integrating multi-dimensional optical encoding technology with AI D²NN, each channel can be dedicated to recognizing only specific object classes while resisting interference from other objects, thus enabling multi-object recognition, as shown in Fig. S19. In addition, by incorporating spatially invariant systems such as optical convolution operators, the AI D²NN can be extended to handle dynamic target recognition based on the convolution results. Besides, the operating wavelength of the network can be expanded to visible or near-infrared spectral range for broader applications. Further optimization such as energy efficiency, optical loss, and fabrication precision will be required for practical implementation, as discussed in Table. S6 in Supplementary Note 9. Furthermore, by capitalizing on the lightweight and planar advantages of metasurface, we can integrate metasurface with the mature complementary metal-oxide semiconductor imaging sensor, to fabricate

an integrated and compact AI D²NN system. Overall, AI D²NN will accelerate their transition from laboratory prototypes to real-world applications, demonstrating strong potential for deployment in autonomous driving perception, precise medical diagnostics, and intelligent security monitoring.

Materials and methods

Numerical calculations

We employed the finite-difference time-domain (FDTD) method to simulate the optical response of microstructures with varying size. The modeled structures were subwavelength cylindrical rods made of high-impedance silicon (refractive index of 3.45), illuminated by a linearly x-polarized plane wave at 0.85 THz. The Si substrate thickness was set to 760 μm , and the height of each rod was fixed at 240 μm . The diameter of the microstructures was varied from 30 μm to 90 μm in 300 nm increments. Periodic boundary conditions were applied in the x and y directions with a unit cell period of 100 μm to model an infinite array. Along the z-direction, a perfectly matched layer (PML) was employed to absorb outgoing waves and eliminate artificial reflections. To characterize the optical response, a point monitor was placed to record the transmitted wave phase, while a planar field positioned above the silicon rods captured the transmission intensity. By analyzing both phase and transmittance, a comprehensive insight into how rod's size interact with the incident terahertz waves were gained.

Training of the AI D²NN

The numerical model of the proposed AI D²NN was deployed on a computer for training. The hardware configuration includes an NVIDIA GeForce RTX 4060 GPU and an Intel® Xeon® Gold 6248 R CPU. The software environment utilized Python 3.8.17 and PyTorch 2.0.1+cu117.

A dataset consisting of 40,000 handwritten digit images and interference images was used for training, with a batch size of 8 and a learning rate of 0.01. During training, the Mean Squared Error (MSE) loss function and the Pearson Correlation Coefficient (PCC) was selected to calculate loss values, these metrics are commonly used for target classification and imaging quality assessment in machine learning task. The Adam optimizer was employed to fine-tune the phase parameters of each layer based on the computed gradients. Above process is repeated many times until the network converged.

Experimental setup

The AI D²NN was experimentally implemented within a terahertz time-domain spectroscopy (THz-TDS) system. Femtosecond laser pulses were split into pump and probe



paths: the pump beam generates terahertz radiation, while the probe beam was used to record the transmitted terahertz waveform after passing through the network. A CCD electro-optic sampling system (MER-531-20GM-P) was employed to scan the terahertz pulses, enabling the Fourier transformation of the recorded time-domain signals to obtain their spectral components in the frequency domain. To produce an ideal input intensity profile, a stainless-steel mask was used to block portions of the terahertz beam, creating a customized spatial intensity distribution. The custom 3D-printed holder (printer model: JG-A240, fabrication accuracy: 100 μm) ensured precise alignment and interlayer spacing between the mask and diffractive layers.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U21A20140, No. 62575023), the National Key Research and Development Program of China (2021YFA1401200), and the Beijing Natural Science Foundation (JQ24028).

Author details

¹Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China. ²National Key Laboratory on Near-surface Detection, Beijing 100072, China. ³Beijing Key Laboratory of Metamaterials and Devices, Key Laboratory of Terahertz Optoelectronics of Ministry of Education, Capital Normal University, Department of Physics, Capital Normal University, Beijing 100048, China. ⁴Qiyuan Lab, Beijing 100095, China. ⁵Engineering Product Development, Singapore University of Technology and Design, Singapore 487372, Singapore. ⁶Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

Author contributions

Conceptualization: Z.H. and N.Z. Methodology: Z.H., Y.L., Z.Z., Q.L., C.H., and Y.Z. Software: Z.H., Z.Z., S.L. Data curation: Z.H., N.Z., and L.H. Writing—original draft: Z.H. Writing—review and editing: N.Z., Y.L., H.W., X.Q., J.Y., and Y.Z. Visualization: Z.H., Y.L., and N.Z. Project administration: N.Z. and L.H. Funding acquisition: N.Z. and L.H.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Conflict of interest

The authors declare no competing interests.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41377-026-02188-7>.

Received: 18 August 2025 Revised: 18 December 2025 Accepted: 9 January 2026

Published online: 03 February 2026

References

- Kalake, L. et al. Analysis based on recent deep learning approaches applied in real-time multi-object tracking: a review. *IEEE Access* **9**, 32650–32671 (2021).
- Sun, S. J. et al. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 104–119 (2021).
- Yan H. et al. MIML-FCN plus : multi-instance multi-label learning via fully convolutional networks with privileged information. in *Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017).
- Yang, X. H. et al. Deep transfer learning-based multi-object detection for plant stomata phenotypic traits intelligent recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **20**, 321–329 (2023).
- Chen, Z. D. et al. High-resolution feature pyramid network for small object detection on drone view. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 475–489 (2024).
- Zheng, H. Y. et al. Multichannel meta-imagers for accelerating machine vision. *Nat. Nanotechnol.* **19**, 471–478 (2024).
- Zaidi, S. S. A. et al. A survey of modern deep learning based object detection models. *Digit. Signal. Process.* **126**, 103514 (2022).
- Kong, L. Y. et al. LocalLock: enhancing multi-object tracking in satellite videos via local feature matching. *Remote Sens.* **17**, 371 (2025).
- Zhu, T. Y. et al. Looking beyond two frames: end-to-end multi-object tracking using spatial and temporal transformers. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 12783–12797 (2023).
- Wen, T. et al. Semantically enhanced multi-object detection and tracking for autonomous vehicles. *IEEE Trans. Rob.* **39**, 4600–4615 (2023).
- Chen, T. I. et al. Dual-awareness attention for few-shot object detection. *IEEE Trans. Multimed.* **25**, 291–301 (2023).
- Zhu, L. C. et al. Label independent memory for semi-supervised few-shot video classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 273–285 (2022).
- Wang, L. M. et al. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 2740–2755 (2019).
- Li, Z. Y. et al. VideoLSTM convolves, attends and flows for action recognition. *Comput. Vis. Image Underst.* **166**, 41–50 (2018).
- Lin, Z. J. et al. 120 GOPS Photonic tensor core in thin-film lithium niobate for inference and in situ training. *Nat. Commun.* **15**, 1–10 (2024).
- Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
- Miller, D. A. B. Attosecond optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
- Lee, S. H. et al. Nanoscale resistive switching devices for memory and computing applications. *Nano Res.* **13**, 1228–1243 (2020).
- Feng, F. et al. Symbiotic evolution of photonics and artificial intelligence: a comprehensive review. *Adv. Photonics* **7**, 024001 (2025).
- Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
- Ashtiani, F. et al. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
- Bai, B. J. et al. All-optical image classification through unknown random diffrusers using a single-pixel diffractive network. *Light Sci. Appl.* **12**, 14 (2023).
- Huang, Z. B. et al. All-optical signal processing of vortex beams with diffractive deep neural networks. *Phys. Rev. Appl.* **15**, 36936–36952 (2021).
- Chen, Z. J. et al. Deep learning with coherent VCSEL neural networks. *Nat. Photonics* **17**, 723–730 (2023).
- Hu, J. T. et al. Diffractive optical computing in free space. *Nat. Commun.* **15**, 1–21 (2024).
- Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
- Zou, X. P. F. et al. Target recognition based on pre-processing in computational ghost imaging with deep learning. *Opt. Laser Technol.* **167**, 1–9 (2023).
- Jiang, Y. et al. Photonic convolution neural network based on interleaved time-wavelength modulation. *J. Lightwave Technol.* **39**, 4592–4600 (2021).
- Li, J. X. et al. Massively parallel universal linear transformations using a wavelength-multiplexed diffractive optical network. *Adv. Photonics* **5**, 016003 (2023).
- Chi, H. X. et al. Metasurface enabled multi-target and multi-wavelength diffraction neural networks. *Laser Photonics Rev.* **19**, 158 (2025).
- Duan, Z. Y. et al. Optical multi-task learning using multi-wavelength diffractive deep neural networks. *Nanophotonics* **12**, 893–903 (2023).
- Xiang, C. X. et al. Multiplexed metasurfaces for diffractive optics via a phase correlation method. *Opt. Lett.* **50**, 1989–1992 (2025).
- Luo, Y. et al. Design of task-specific optical systems using broadband diffractive neural networks. *Light Sci. Appl.* **8**, 112 (2019).
- Li, J. X. et al. Polarization multiplexed diffractive computing: all-optical implementation of a group of linear transformations through a polarization-encoded diffractive network. *Light Sci. Appl.* **11**, 153 (2022).
- Luo, X. H. et al. Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible. *Light Sci. Appl.* **11**, 158 (2022).
- Wang, Y. Z. et al. Matrix diffractive deep neural networks merging polarization into meta-devices. *Laser Photonics Rev.* **18**, 2300903 (2024).



37. Li, B. L. et al. Orbital angular momentum optical communications enhanced by artificial intelligence. *J. Opt.* **24**, 094003 (2022).
38. Fang, X. Y. et al. Orbital angular momentum-mediated machine learning for high-accuracy mode-feature encoding. *Light Sci. Appl.* **13**, 49 (2024).
39. Cheng, J. W. et al. Multimodal deep learning using on-chip diffractive optics with in situ training capability. *Nat. Commun.* **15**, 6189 (2024).
40. Wei, M. L. et al. Electrically programmable phase-change photonic memory for optical neural networks with nanoseconds in situ training capability. *Adv. Photonics* **5**, 046004 (2023).
41. Xu, S. F. et al. High-order tensor flow processing using integrated photonic circuits. *Nat. Commun.* **13**, 7970 (2022).
42. Tait, A. N. et al. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017).
43. Wu, C. M. et al. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nat. Commun.* **12**, 96 (2021).
44. Guarino, A. et al. Electro-optically tunable microring resonators in lithium niobate. *Nat. Photonics* **1**, 407–410 (2007).
45. Zhu, D. et al. Integrated photonics on thin-film lithium niobate. *Adv. Opt. Photonics* **13**, 242–352 (2021).
46. Chen, Y. T. et al. All-analog photoelectronic chip for high-speed vision tasks. *Nature* **623**, 48–57 (2023).
47. Dong, B. et al. Partial coherence enhances parallelized photonic computing. *Nature* **632**, 55–62 (2024).
48. Qin, J. et al. All-optical Fourier neural network using partially coherent light. *Chip* **4**, 100140 (2025).
49. He, C. et al. Pluggable multitask diffractive neural networks based on cascaded metasurfaces. *Opto-Electron. Adv.* **7**, 230005 (2024).
50. Bai, B. J. et al. Pyramid diffractive optical networks for unidirectional image magnification and demagnification. *Light Sci. Appl.* **13**, 178 (2024).
51. Bai, B. J. et al. To image, or not to image: class-specific diffractive cameras with all-optical erasure of undesired objects. *Elight* **2**, 14 (2022).
52. Ma, G. D. et al. Unidirectional imaging with partially coherent light. *Adv. Photonics Nexus* **3**, 066008 (2024).
53. Xu, Z. H. et al. Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. *Science* **384**, 202–209 (2024).
54. Zhang, H. et al. An optical neural chip for implementing complex-valued neural network. *Nat. Commun.* **12**, 457 (2021).
55. Xue, Z. W. et al. Fully forward mode training for optical neural networks. *Nature* **632**, 280–286 (2024).
56. Li, J. X. et al. Class-specific differential detection in diffractive optical neural networks improves inference accuracy. *Adv. Photonics* **1**, 046001 (2019).
57. Zhu, H. H. et al. Space-efficient optical computing with an integrated chip diffractive neural network. *Nat. Commun.* **13**, 1044 (2022).
58. Qian, C. et al. Dynamic recognition and mirage using neuro-metamaterials. *Nat. Commun.* **13**, 2694 (2022).
59. Huang, Z. et al. Pre-sensor computing with compact multilayer optical neural network. *Sci. Adv.* **10**, eado8516 (2024).

