Opto-Electronic Advances

Supplementary information

February 2024, Vol. 7, No. 2

DOI: 10.29026/oea.2024.230005

Pluggable multitask diffractive neural networks based on cascaded metasurfaces

Cong He[1], Dan Zhao[2], Fei Fan[2], Hongqiang Zhou[1,3], Xin Li[1], Yao Li[4], Junjie Li[4], Fei Dong[5], Yin-Xiao Miao[5], Yongtian Wang[1]* and Lingling Huang[1]*

[1]Beijing Engineering Research Center of Mixed Reality and Advanced Display, Key Laboratory of Photoelectronic Imaging Technology and System of Ministry of Education of China, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China; [2]Institute of Modern Optics, Tianjin Key Laboratory of Optoelectronic Sensor and Sensing Network Technology, Nankai University, Tianjin 300350, China; [3]Department of Physics and Optoelectronics, Faculty of Science, Beijing University of Technology, Beijing 100124, China; [4]Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100191, China; [5]Beijing Aerospace Institute for Metrology and Measurement Technology, Beijing 100076, China.

*Correspondence: YT Wang, E-mail: wyt@bit.edu.cn; Huang LL, E-mail: huanglingling@bit.edu.cn

This file includes:

Supplementary information for this paper is available at https://doi.org/10.29026/oea.2024.230005

Section 1: Error Backpropagation

We use the error back propagation algorithm and the random gradient descent optimization method. The algorithm defines a mean square error (MSE) loss function to evaluate the performance of P-DNN output for the desired target. Assuming that the Nth layer is the last layer of P-DNN, the intensity of optical field measured by the detector on the output plane is $s_i^{N+1} = \left|U_i^{N+1}\right|^2$, i refers to a neuron. The error loss between the output plane $s_i^{N+1}$and the target g_i^{N+1} can be expressed as:

$$F\left(t_i^l\right) = \frac{1}{K}\sum_k \left(s_k^{N+1} - g_k^{N+1}\right)^2 , \tag{S1}$$

where K refers to the number of measurement points at the output plane. The process of network parameter optimization can be seen as the process of minimizing function $F\left(t_i^l\right)$. When the phase only modulation method is selected, the amplitude $a_i{}^l$ is set to 1. The gradient of $F\left(t_i^l\right)$ with respect to $\varphi_i{}^l$ at a given layer l is expressed as:

$$\frac{\partial F\left(\varphi_i^l\right)}{\partial \varphi_i^l} = \frac{4}{K}\sum_k \left(s_k^{N+1} - g_k^{N+1}\right) \cdot Real\left\{\left(u_k^{N+1}\right)^* \cdot \frac{\partial u_k^{N+1}}{\partial \varphi_i^l}\right\} , \tag{S2}$$

u_k^{N+1} is the discrete summation of $U\left(\boldsymbol{r}^l\right)$ in Eq. (2), u_k^{N+1} can be expressed as $u_k^{N+1} = \sum_{k1} h_{k1}^N\left(x_k, y_k, z_k\right) \cdot t_{k1}^N\left(x_{k1}, y_{k1}, z_{k1}\right) \cdot u_{k1}^N\left(x_{k1}, y_{k1}, z_{k1}\right)$. Then, $\frac{\partial u_k^{N+1}}{\partial \varphi_i^l}$ can be expressed as:

$$\frac{\partial u_k^{N+1}}{\partial \varphi_i^{l=N}} = \mathrm{j} \cdot t_i^N\left(x_i, y_i, z_i\right) \cdot u_i^N\left(x_i, y_i, z_i\right) \cdot h_i^N\left(x_k, y_k, z_k\right) , \tag{S3}$$

For every layer, $l \le N$, this gradient can be calculated using:

$$\frac{\partial u_k^{N+1}}{\partial \varphi_i^{l=N-1}} = \mathrm{j} \cdot t_i^{N-1} \cdot u_i^{N-1}\left(x_i, y_i, z_i\right) \cdot \sum_{k_1} h_{k_1,k}^N\left(x_{k1}, y_{k1}, z_{k1}\right) \cdot t_{k_1}^N\left(x_{k1}, y_{k1}, z_{k1}\right) \cdot h_{i,k_1}^{N-1}\left(x_k, y_k, z_k\right) , \tag{S4}$$

$$\frac{\partial u_k^{N+1}}{\partial \phi_i^{l=N-2}} = \mathrm{j} \cdot t_i^{N-2} \cdot u_i^{N-2} \cdot \sum_{k_1} h_{k_1,k}^N \cdot t_{k_1}^N \cdot \sum_{k_2} h_{k_2,k_1}^{N-1} \cdot t_{k_2}^{N-1} \cdot h_{i,k_2}^{N-2} , \tag{S5}$$

$$\cdots$$

$$\frac{\partial u_k^{N+1}}{\partial \varphi_i^{l=N-L}} = \mathrm{j} \cdot t_i^{N-L} \cdot u_i^{N-L} \cdot \sum_{k_1} h_{k_1,k}^N \cdot t_{k_1}^N \cdot\cdot\cdot\cdot\cdot \sum_{k_L} h_{k_L,k_{L-1}}^{N-L+1} \cdot t_{k_L}^{N-L+1} \cdot h_{i,k_L}^{N-L} , \tag{S6}$$

where $2 \le L \le N-1$. In each iteration, the training data is input into P-DNN to obtain the loss function, which is then used to update the entire network parameters.

Section 2: Simulation and experimental results

Similar to Fig. 5, 6 in the text, Fig. S1 shows the complete test results. The output results obtained by taking handwritten numerals and fashion datasets as input are shown. According to the energy distribution of the output plane, it can be concluded that P-DNN can successfully identify multiple tasks, while ensuring high accuracy.

Section 3: Analysis of alignment error between network layers

We conducted a theoretical analysis to understand the impact caused by this alignment error. The identification accuracy of P-DNN is affected by the displacement along x, y and z direction between metasurfaces. The errors caused by displacement along the x direction and y direction are similar. Figure S2(a) shows the simulation results obtained by the 1-pix (5 μm) offset of the two layers of metasurfaces in the x direction, and the recognition accuracy is reduced from 92.8% to 58.5%. Similarly, we also analyzed the errors caused by z direction. When errors of z direction exceeds 100 μm, the recognition accuracy decreases with a small decrease from 92.8% to 90.7% (Fig. S2(b, c)). It can be inferred that to ensure the identification accuracy of the diffractive neural network, it is necessary to ensure that the errors of the x direction and y direction is less than 1-pix and the errors of z direction is less than 100 μm.

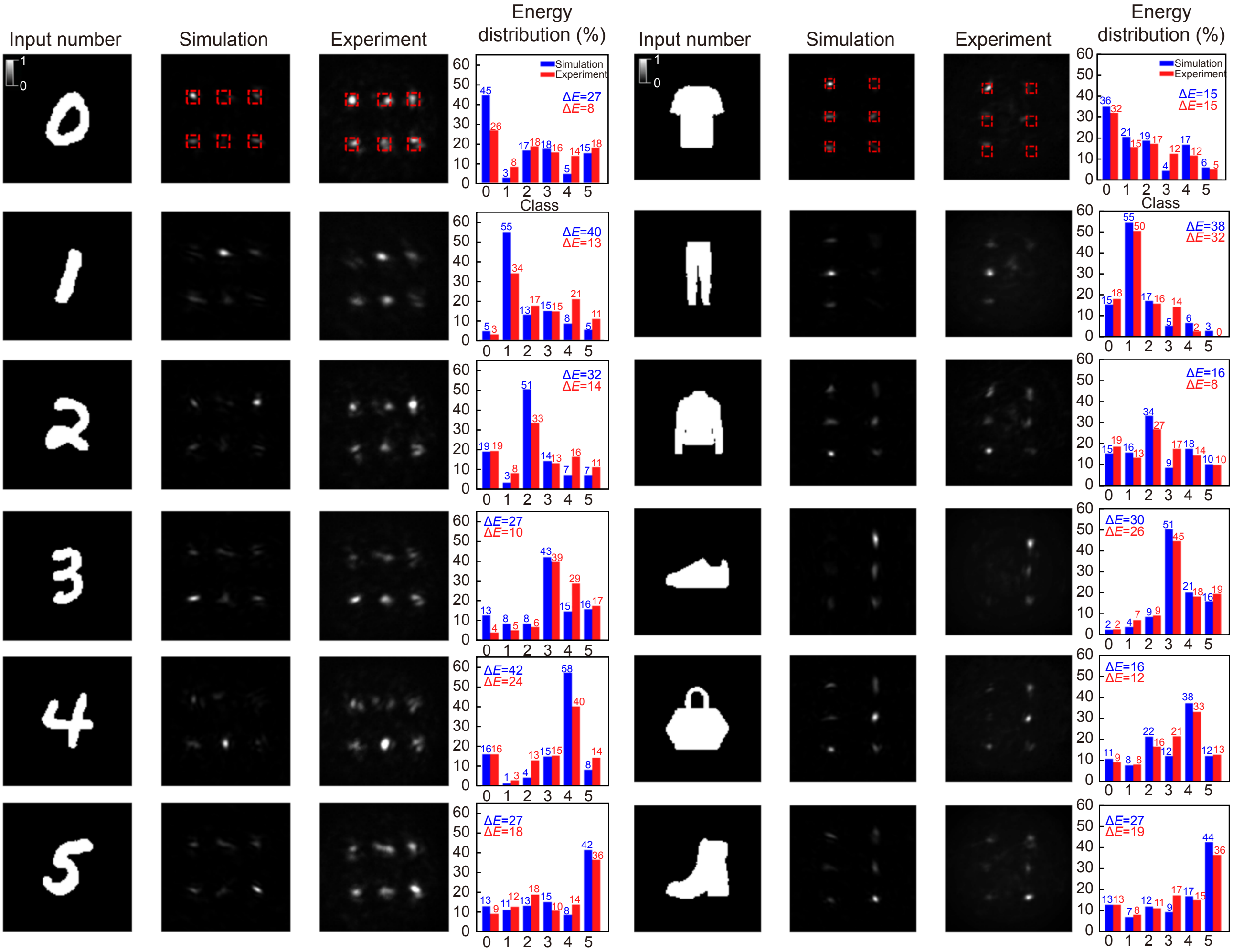


Fig. S1 | Simulation and experimental results of handwritten digital and fashion P-DNN. The percentage of energy distribution corresponding to each input data on the output plane shows that P-DNN accurately focuses the energy to the preset area, and can accurately identify the input object according to the energy distribution. Δ*E* represents the difference between the percentage of maximum and second maximum energy.

Section 4: The polarization conversion efficiency

We have measured the transmission of our metasurfaces at multiple wavelengths, and then calculate its polarization conversion efficiency by using $\eta=\frac{T_l}{I_r}$, where I_r is the intensity of the incident left circularly polarized light and T_l is intensity of the cross circularly polarized light through the single layer metasurfaces. The measured polarization conversion efficiency of single layer metasurfaces is shown in Fig. S3(a). Note the measured polarization conversion can reach up to 82% at around 800 nm, while at other wavelengths between 740 nm and 900 nm, the efficiencies are also above 50%. For cascaded metasurfaces, because the polarization should be converted twice, the polarization conversion efficiency is defined by $\eta=\frac{T_l^1\cdot T_r^2}{I_r}$, where T_l^1 and T_r^2 are the cross circularly polarized transmitted coefficient of the first and second metasurfaces, respectively, and corresponding experimental result is about 51.3% at 800 nm (Fig. S3(b)). The conversion efficiency of cascaded metasurfaces is lower than the product of conversion efficiency of single layer metasurfaces, because there is scattering light and other losses between layers.To better demonstrate the recognition performance, we simulated the process by taking 80% of the energy of the outgoing light at the first layer and adding 20% of the unmodulated incident energy, and do the same processing at the second layer. As shown in Fig. S4, it can be seen that there is only a small energy deviation from the normal simulation, which does not affect the testing accuracy. Moreover, we can see from the experimental section in the manuscript that the unmodulated light does not affect the recognition accuracy of the objects.

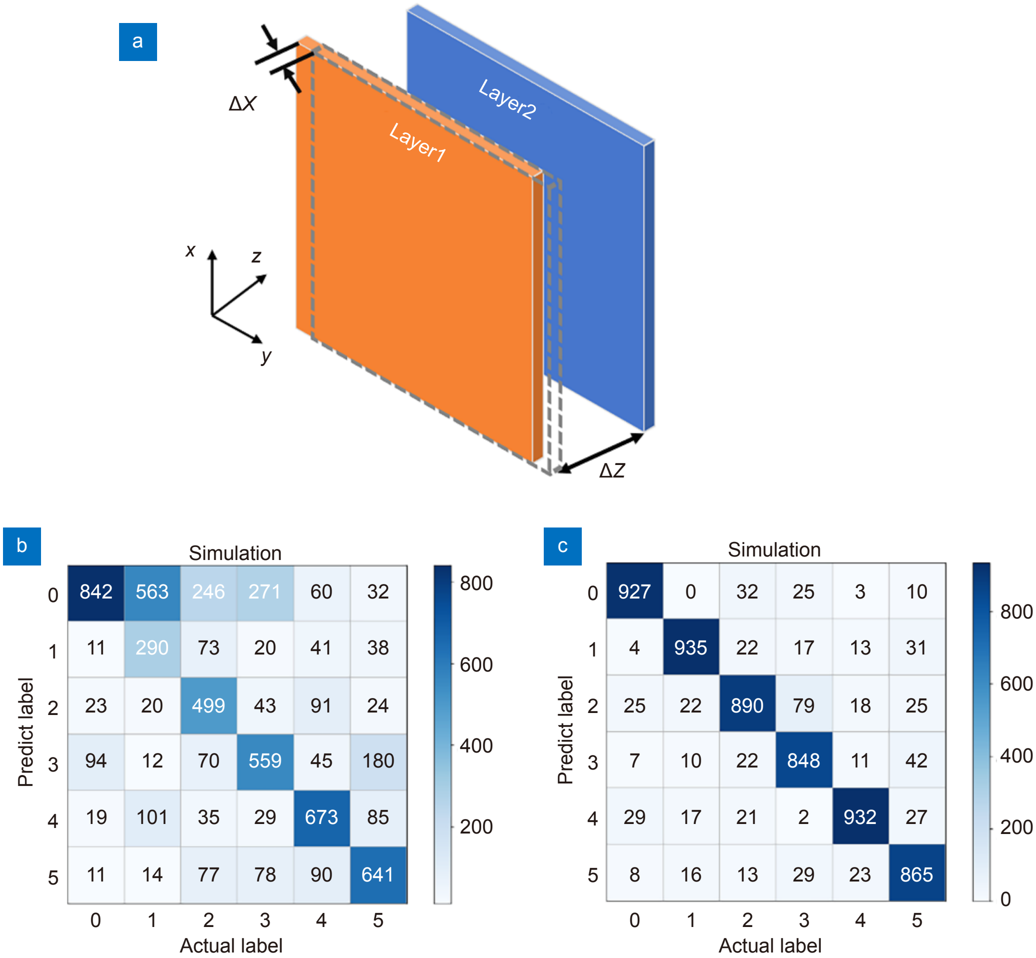


Fig. S2 | Analysis of alignment error between network layers. (**a**) Schematic diagram of network alignment, where Δ*x* represents errors of the *x* direction and Δ*z* represents errors of *z* direction (**b**) Confusion matrix corresponding to simulation results when errors of the *x* direction are 1pix (5 μm). (**c**) Confusion matrix corresponding to simulation results when errors of *z* direction is 100 μm.

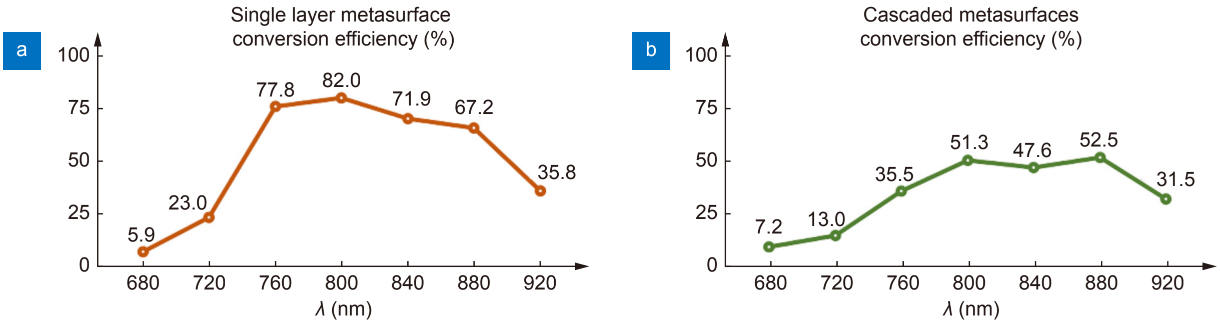


Fig. S3 | The measured polarization conversion efficiency. (**a**) The conversion efficiency of single layer metasurfaces. (**b**) The conversion efficiency of cascaded metasurfaces.

Section 5: Factors affecting the performance of P-DNN

The depth of a P-DNN has a significant impact on its ability to perform complex tasks, and there is a limit to the performance when the network depth is fixed. Lin et al. have demonstrated that increasing the depth of the network can provide greater training freedom and effectively improve the performance of the network[S1]. Therefore, there are two ways to enhance the performance of P-DNN. The most effective way is to increase the depth of the network, which can effectively improve the performance of the network and be used for more complex tasks. The other way is to increase the pixel numbers of each layer, which can also improve the performance of the network to a certain extent. When us-

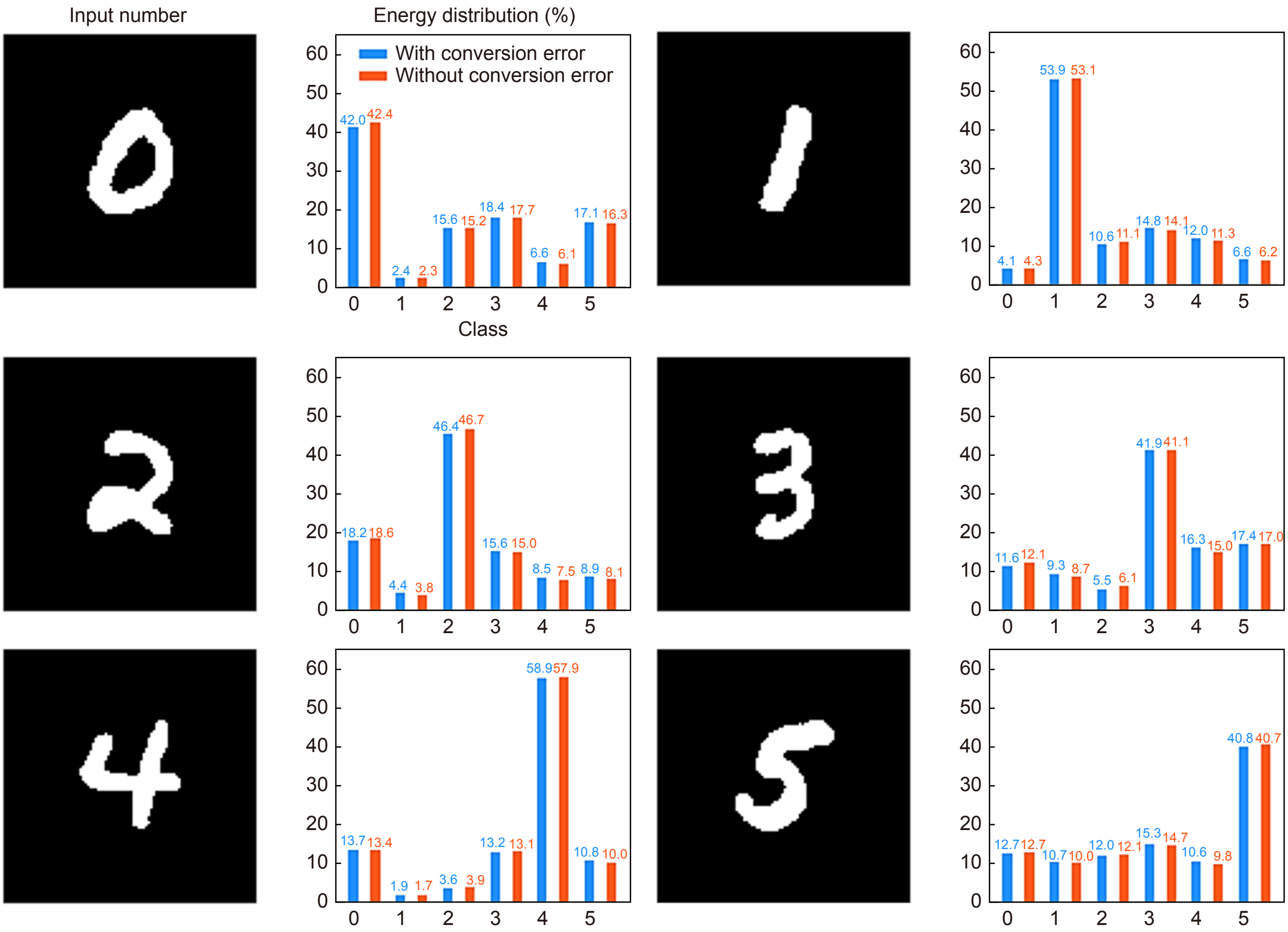
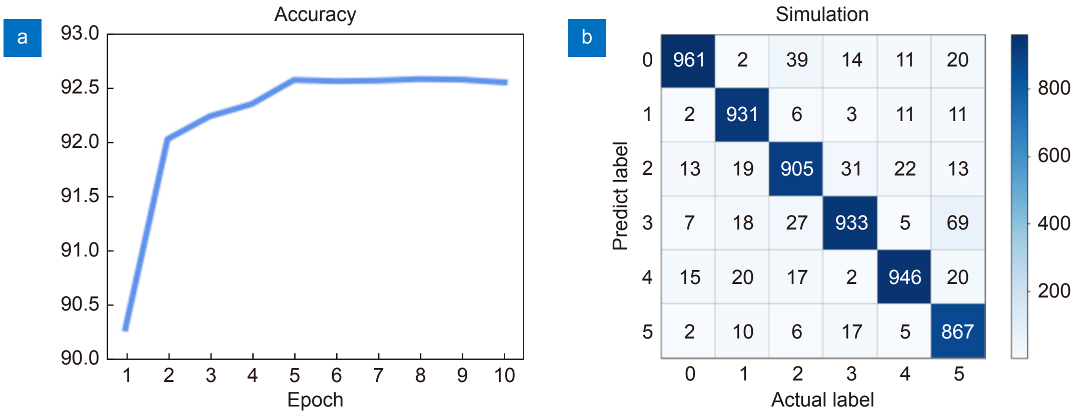


Fig. S4 | A comparison is made between the energy distribution obtained from unmodulated light and the ideal case. The energy distributions of digits "0" to "5" are presented, respectively, showing that the impact of unmodulated light on recognizing energy distribution is negligible.

ing a 2-layer P-DNN with pixel numbers of 200×200 (Fig. S5), the test accuracy on the "0"-"5" dataset reached 92.6%, which is a 0.8% improvement compared to the results reported (91.8%) in the manuscript. When the depth and pixel numbers of the network increases, P-DNN can be used to perform more complex tasks such as classification with more categories, feature extraction, etc.



	0	1	2	3	4	5
0	961	2	39	14	11	20
1	2	931	6	3	11	11
2	13	19	905	31	22	13
3	7	18	27	933	5	69
4	15	20	17	2	946	20
5	2	10	6	17	5	867

Fig. S5 | Simulation result of the 200-pixel size of the P-DNN. (**a**) Test accuracy is 92.6% after training 10 epochs. (**b**) The simulation results of the confusion matrix for handwritten digits were obtained by dividing the sum of elements on the main diagonal of the confusion matrix by the total sum of all elements.

Section 6: Additional phase error introduced by DMD

The principle of DMD is to control the reflection angle through rotating micromirrors to achieve amplitude modulation. Therefore, slight variations in optical path differences may occur at different reflection angles, leading to the introduction of phase noise. To analyze the impact of additional phase errors on network testing, we added random phase noise within the range of [−1, 1] rad to the input images during simulation propagation. The test dataset used in this test consisted of the same handwritten digits “0” to “5” as in the manuscript. The energy distribution characteristics in the detection area were obtained for each input image, as shown in Fig. S6. The energy error of each class was within 0.2% for all cases when the noise was added to the input images, which could be considered negligible. The energy was concentrated entirely in the target region, indicating that the phase noise introduced by DMD can be ignored while maintaining a high accuracy rate. To reduce the phase error caused by DMD, one can add some random phase noise during the training process, which can increase the robustness of the training results.

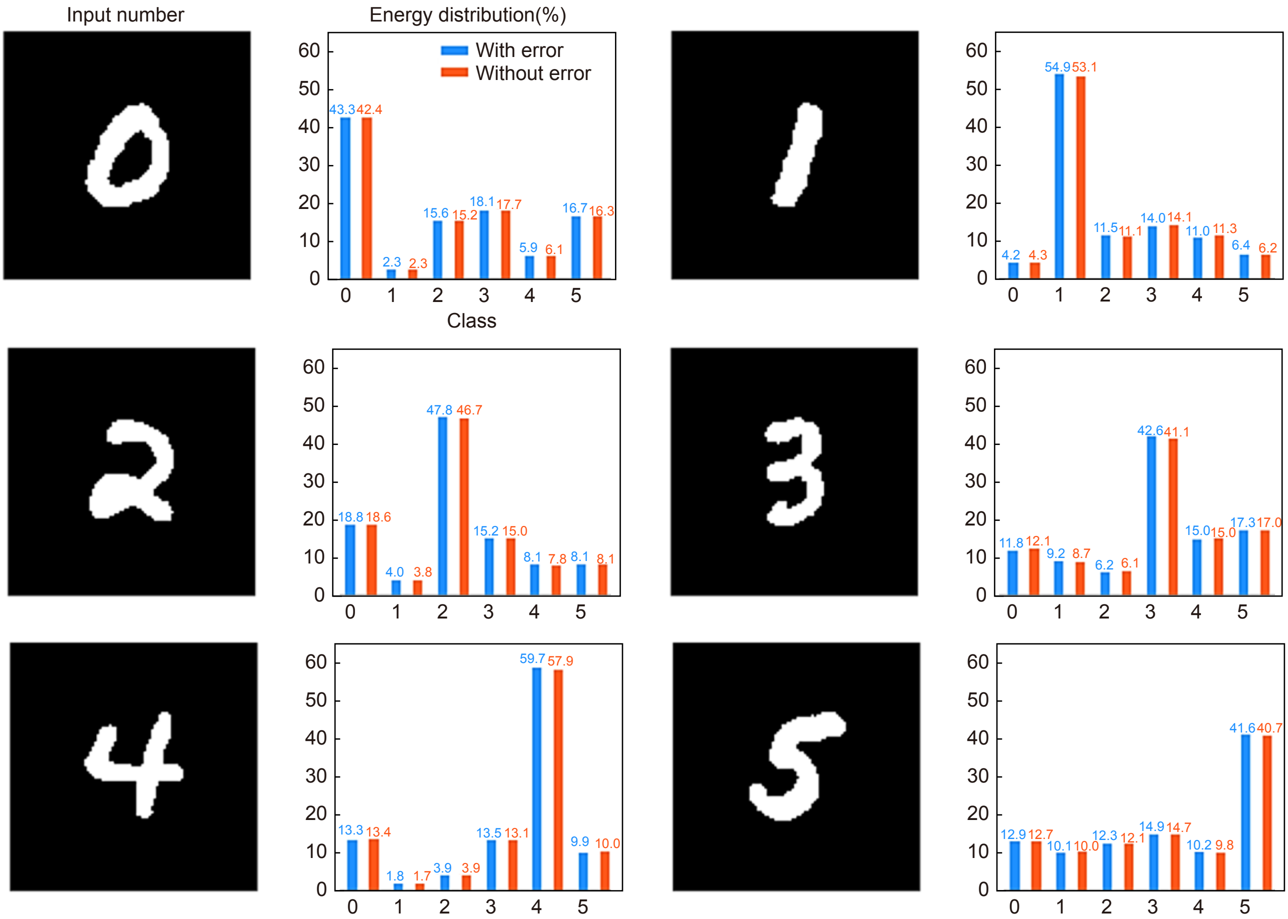


Fig. S6 | The energy distribution with DMD phase error is compared with the ideal case. The energy distributions of digits "0" to "5" are presented separately, showing that the influence of DMD on recognizing energy distribution is negligible.

Section 7: Effect of different numbers of shared layers and classification layers on P-DNN performance

In order to better demonstrate the performance improvement of P-DNN, we used the same six digits and six fashion items as training datasets as in the manuscript. As shown in Fig. S7(a, b), it can be seen that with the increase of layers, P-DNN can exhibit stronger performance and there is a significant increase in recognition accuracy on both datasets. It has been observed that the number of classification layers has a certain degree of influence on the accuracy of task recognition. In particular, for the same task, the 3-layer P-DNN with two classification layers achieves slightly higher accuracy than a 4-layer P-DNN with only one classification layer, but lower accuracy than a 4-layer P-DNN with two classification layers. It also demonstrates that task switching can be achieved by replacing a relatively small number of layers while maintaining high recognition accuracy.

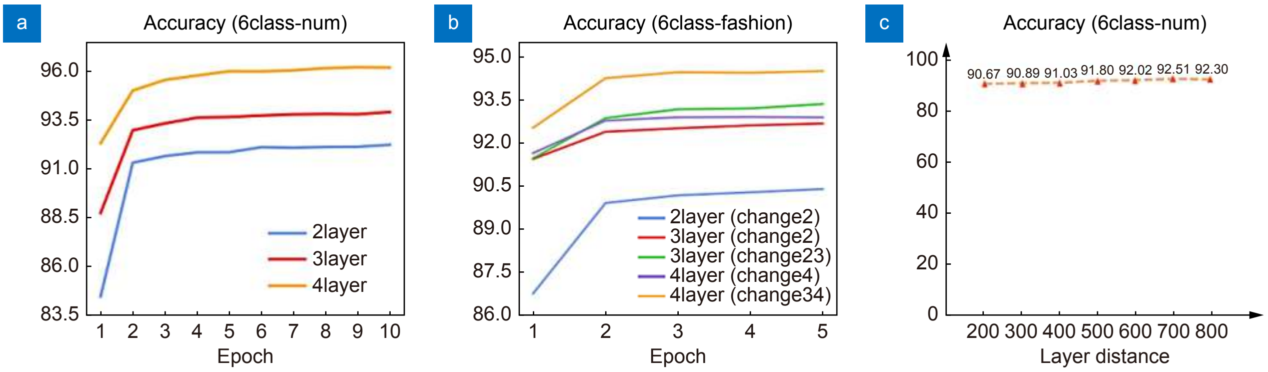


Fig. S7 | Performance of different sharing layers and classification layers and different layers in the six classification tasks. (**a**) Test accuracy of P-DNN with different numbers of layers on the digit dataset. (**b**) Test accuracy of P-DNN with different shared and classification layers on the fashion dataset, where "change2" indicates switching the second layer plug-in. (**c**) Test accuracy of double-layer P-DNN with different separations.

To analyze the performance when adjusting the meta-atom distances and layer distces, we took the 2-layer P-DNN as an example and used the same images of handwritten digits "0" to "5" as test data. We tested the design of P-DNN with different separations and obtained recognition accuracy in the range of 100 μm to 800 μm. The results in Fig. S7(c) show that the recognition accuracy slightly increases as the layer distances gradually increase. Because when the layer distances are small, the network connectivity will be reduced, which affects the recognition accuracy[S2]. Ultimately, we chose 500 μm as the layer distances to balance the alignment difficulty and recognition accuracy in our manuscript.

Subsequently, we have attempted to implement more complex classification tasks using three and four layers in simulation. As shown in Fig. S8, it can be observed that using a 3-layer P-DNN can achieve recognition of the the whole MNIST and Fashion-MNIST datasets, but if higher accuracy is required, the number of classification layers should be increased or four layers of P-DNN should be chosen. Similar to the result of 6 classification, the accuracy of 3-layer P-DNN with two classification layers is slightly higher than that of 4-layer P-DNN with single classification layer. Increasing the numbers of layers in the network can improve the recognition accuracy of the system, but the depth of the network is still the most important factor. Within a certain range, increasing the depth of a neural network can enable it to handle more complex tasks. In addition, we also added a training result of a 10-layer network to demonstrate that deeper networks do not necessarily lead to better performance. From Fig. S8(a), it can be observed that compared to the 4-layer network, there is almost no improvement in recognition accuracy, but longer training times and larger computational resources are required. Combining polarization multiplexing or multi-wavelength multiplexing channels of metasurfaces with plugins enable more parallel tasks to be performed, which may be a researchable direction.

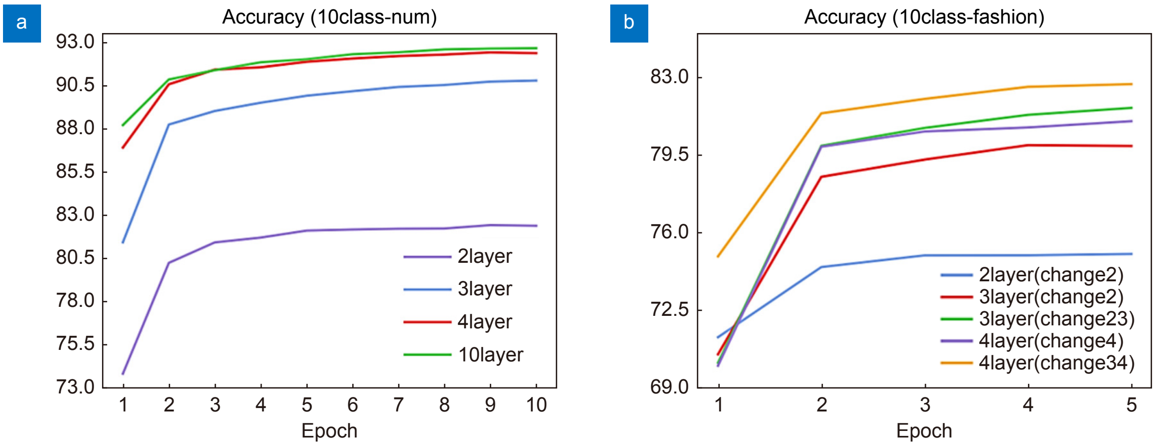


Fig. S8 | Performance of different sharing layers and classification layers and different layers in the ten classification tasks. (**a**) Test accuracy of P-DNN with different numbers of layers on the whole MNIST dataset. (**b**) Test accuracy of P-DNN with different shared and classification layers on the whole Fashion-MNIST dataset.

In addition, the selection of layers is related to task difficulty and information storage limit of each layer. Unfortunately, there is currently no universal formula that can accurately provide the number of layers needed for training based on the complexity of the task. Typically, the number of layers and pixels is chosen based on experience and multiple attempts. In practical application, it is hoped that the number of switchable layers should be as few as possible to complete the task, in order to reduce the difficulties in making pluggable devices in the future.

References

S1. Lin X, Rivenson Y, Yardimci NT, Veli M, Luo Y et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).

S2. Chen H, Feng JN, Jiang MW, Wang YQ, Lin J et al. Diffractive deep neural networks at visible wavelengths. *Engineering* **7**, 1483–1491 (2021).