Supplemental Material of

Hybrid structure for enhancing robustness in optical diffractive neural networks without vaccination training

Haodong Zhu[†], Ruiqi Yin[†], Rui Xia, Minglong Li, Zhengyu Chen, Zhenyu Yang, Ming Zhao[*]

Nanophotonics Laboratory, School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China

(† These authors contributed equally to this work.)

(*Electronic mail: zhaoming@mail.hust.edu.cn)

S1. Free space optical transmission method

For a n-layer visible-light optical diffractive neural network, the field transmittance $T_l(x,y)$ ($l=1,2,...,n$)of each diffraction layer is used to modulate the coherent wave field $U_l(x,y)$ before the diffraction layer. The modulated coherent wave field $U_l^{'}(x,y)=U_l(x,y)T_l(x,y)$ is then propagated to the next diffraction layer at a distance of z using the angular spectrum diffraction:

$$\begin{cases} U_{l+1}(x,y)=\mathcal{F}^{-1}\{\mathcal{F}[U_l^{'}(x,y)]H(f_x,f_y,z)\}, \\ H(f_x,f_y,z)=\mathcal{F}\{\dfrac{z}{j\lambda(x^2+y^2+z^2)}\exp(j\dfrac{2\pi\sqrt{x^2+y^2+z^2}}{\lambda})\}. \end{cases} \tag{S1}$$

where $H(f_x,f_y,z)$ is the angular spectrum diffraction propagation function, $\mathcal{F}$ represents the Fourier transform.

S2. Network construction

In all three tasks, the size of each diffraction layer is 100λ×100λ. The sampling period of the propagation model are chosen as 1λ. ODNNs are all performed using Python (v3.8.15) and TensorFlow (v2.1.0, Google Inc.) on a PC (Windows 10 operating system) with the hardware: central processing unit (CPU), AMD Ryzen 5 5600 6-Core Processor; random access memory (RAM), 16 GB; graphical processing unit (GPU), GeForce RTX 4070. We train the network using the Adam optimizer provided by TensorFlow, with parameters $\beta_1=0.9$, $\beta_2=0.99$, and $\varepsilon=1e^{-7}$.

S2.1 Digital recognition

We use 20000 randomly selected images (digits 0 - 5) from MNIST as the dataset. We simultaneously select 6000 images (digits 0 - 5)from MNIST as the testing set. For each network structure, we train it for a total of 50 epochs using a learning rate of 0.01.

By balancing accuracy and practical simulation effects, a novel loss function is designed as:

$$loss=\frac{\sum_{i=0}^{5}I_i+k\times I_{background}}{I_i}, \tag{S2}$$

where I_i is the intensity of the *i*-th detector area on the output plane, $\sum_{i=0}^{5}I_i$ is the total intensity of the six areas, and $I_{background}$ is the background intensity outside the six areas. The coefficient k is a tuning factor used to adjust the optimization weight of $I_{background}$. Through simulation validation, k=0.1 was determined. This loss function not only optimizes the network's accuracy but also considers the background light intensity, allowing the output plane to have better light intensity contrast, which is beneficial for achieving improved experimental results.

S2.2 Image denoising

We use 20000 randomly selected images (digits 0 - 9) from MNIST as the dataset. We simultaneously select 6000 images (digits 0 - 9)from MNIST as the testing set. For each network structure, we train it for a total of 50 epochs using a learning rate of 0.01. Mean square error(MSE) was used as the loss function to train the network:

$$MSE=\frac{1}{m\times m}\sum_{i=1}^{m\times m}(Y_i-\hat{Y}_i)^2 \tag{S3}$$

where Y_i is the target light intensity distribution matrix, $\hat{Y}_i$ is the light intensity distribution matrix of the output plane, *i* represents the *i*-th neuron.

S2.3 Wavelength classification

In this task, we use plane waves with wavelengths of 0.9λ and 1.1λ as inputs, the height distribution of the phase plate serves as the variable for random optimization. For each network structure, we train it for a total of 2000 epochs using a learning rate of (0.01×λ/π).

The loss function for training is defined as the average of the loss functions at the two wavelengths, for each wavelength, MSE was used as the loss function to train the network:

$$MSE = \frac{1}{m \times m} \sum_{i=1}^{m \times m} (Y_i - \hat{Y}_i)^2 \tag{S4}$$

where Y_i is the target light intensity distribution matrix, $\hat{Y}_i$ is the light intensity distribution matrix of the output plane，i represents the i-th neuron.

S3. Testing for rotation and scaling errors

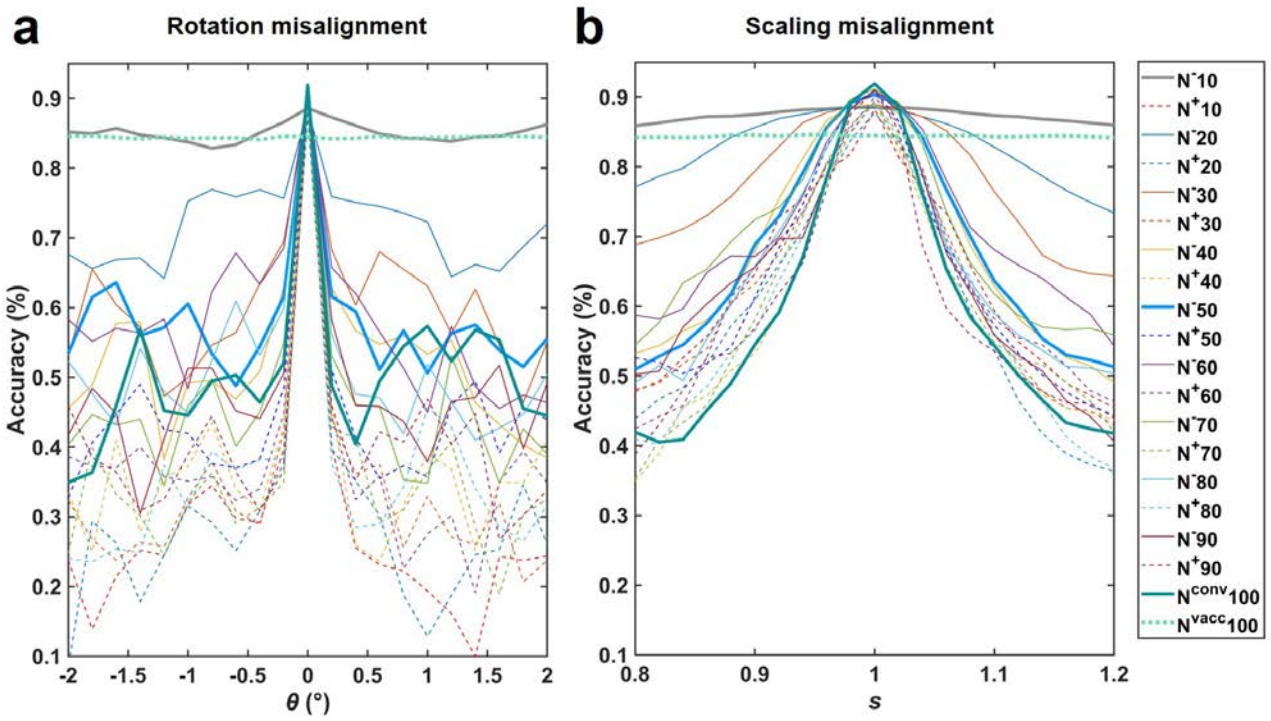

Fig. S1 Testing for rotation and scaling errors in digit recognition task. **a** Rotation misalignment. **b** Scaling misalignment.

As shown in **Fig. S1**, for the digit recognition task in the **main text**, when the 4 selected networks （$N^{-}10$，$N^{-}50$，$N^{conv}100$，$N^{vacc}100$）under ideal condition, the classification accuracy are 88.58%, 90.33%, 91.92%, and 84.48%, respectively; Under rotate misalignments(-2°, 2°), the accuracy declines by an average of 3.69%, 34.10%, 43.56%, and 0.14%, respectively; Under scaling(0.8, 1.2) misalignments, the accuracy declines by an average of 1.26%, 23.91%, 35.29%, and 0.14%, respectively.

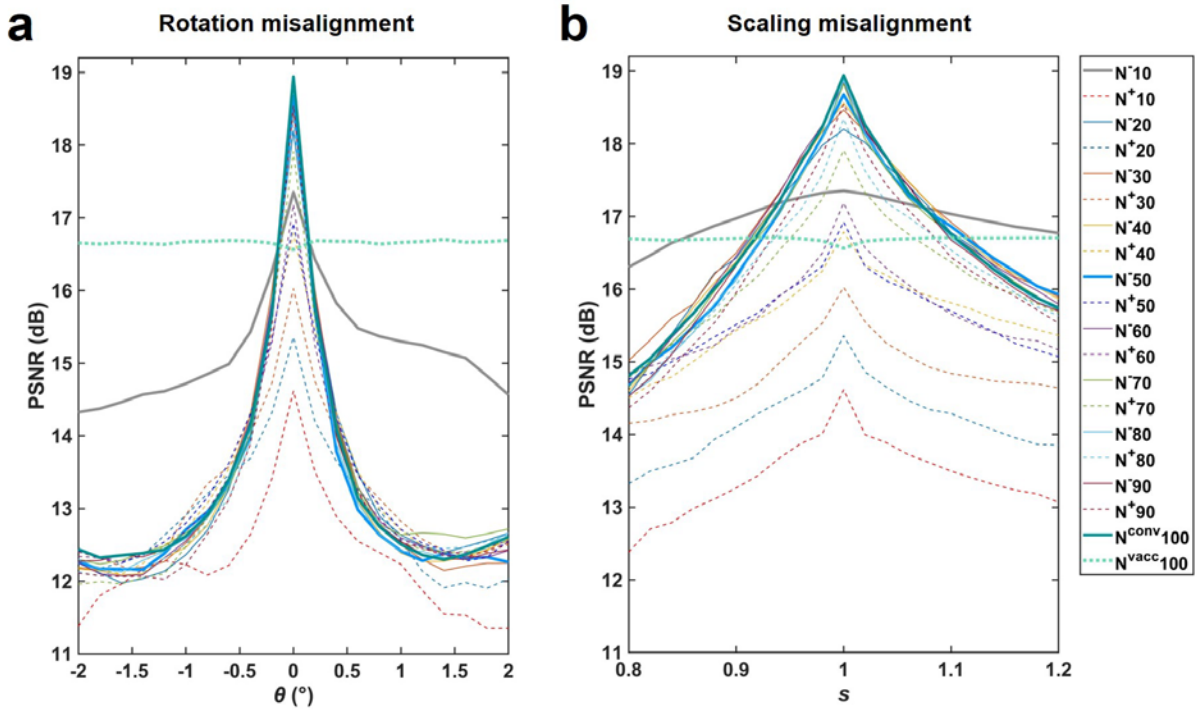

Fig. S2 Testing for rotation and scaling errors in image denoising task. **a** Rotation misalignment. **b** Scaling misalignment.

As shown in **Fig. S2**, for the image denoising task in the **main text**, when the 4 selected networks under ideal condition,the average PSNR are17.39dB, 18.86dB, 19.18dB, and 16.56dB, respectively; Under rotate misalignments, the average PSNR declines by an average of 2.26dB, 5.73dB, 5.89dB, and 0.11dB, respectively;Under scaling misalignments, the average PSNR declines by an average of 0.40dB, 2.18dB, 2.40dB, and 0.13dB, respectively.

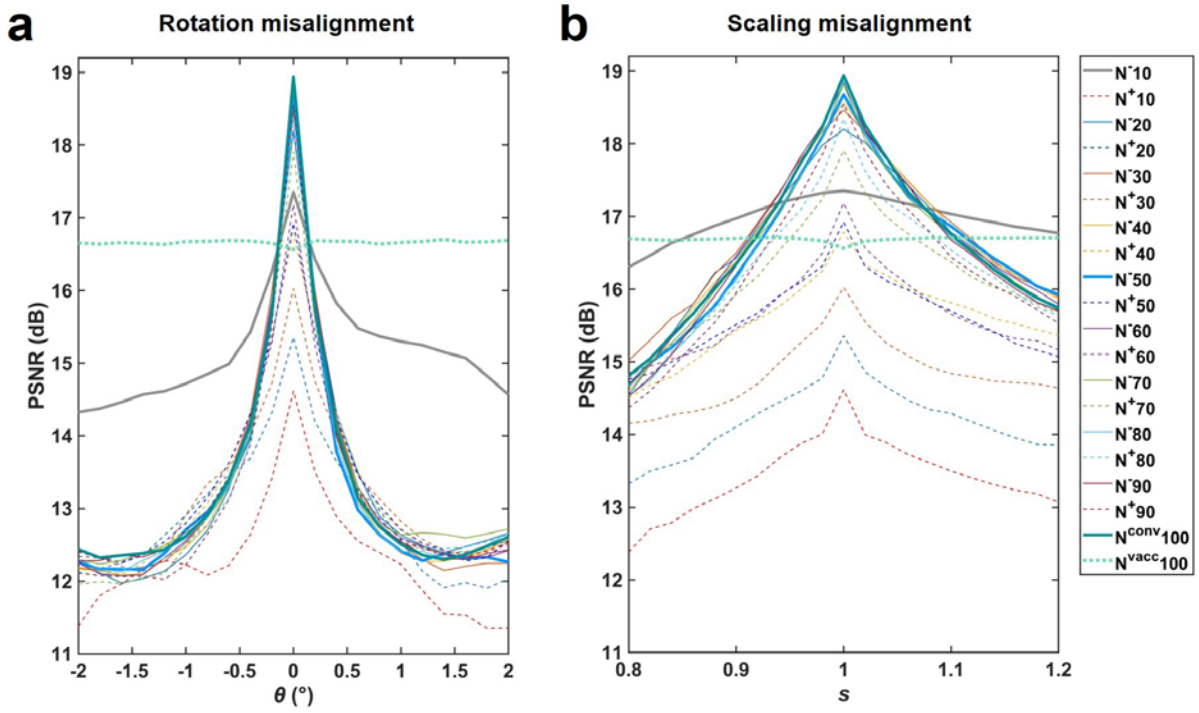

Fig. S3 Testing for rotation and scaling errors in wavelength classification task. **a** Rotation misalignment. **b** Scaling misalignment.

As shown in **Fig. S3**, for the wavelength classification task in the **main text**, when the 4 selected networks under ideal condition, P_{λ_1} are 91.32%, 94.75%, 94.01%, and 52.76%, respectively; P_{λ_2} are 88.27%, 93.34%, 94.11%, and 48.62%, respectively; Under rotate misalignments, the P_{λ_1} declines by an average of 10.84%, 44.23%, 45.72%, and 5.25%, respectively; the P_{λ_2} declines by an average of 16.03%, 46.68%, 45.11%, and 2.52%, respectively;Under scaling misalignments, the P_{λ_1} declines by an average of 4.52%, 35.52%, 42.35%, and 5.30%, respectively; the P_{λ_2} declines by an average of 6.14%, 37.16%, 41.00%, and 3.17%, respectively.

As can be observed, the H-ODNN maintains a certain degree of robustness against rotation and scaling errors. It is worth noting that in the wavelength classification task, the H-ODNNs exhibit relatively poor robustness to rotation errors. This is because introducing rotation errors involves interpolation, which may lead to significant deviations in the optical field computations. Moreover, because the network trained with inoculation must account for more complex error scenarios compared to the two-error case described in the **main text**, its baseline performance under error-free conditions shows some degradation. Moreover, the training time for the three tasks increased by 78%, 240%, and 21%, respectively. In practical applications, thanks to high-precision micro-/nano-fabrication techniques, scaling errors in the phase plate dimensions can be almost neglected. In addition, inter-layer rotation errors and the like can be avoided through multi-point alignment methods. In summary, the H-ODNN proposed in this paper demonstrates robustness under multiple error scenarios and, compared to conventional error-inoculation training, offers better baseline performance and faster training speed.

S4. Comparison of the performance of vaccinated networks with different error vaccination ranges.

In the **main text**, the error-inoculation range used during training of the error-inoculated network is consistent with the range applied during testing. To investigate the impact of different error-inoculation ranges on the performance of the inoculated network, we conducted additional simulations using an inoculation range reduced to half of the test range. The simulation results are shown in **Fig. S4**, with the corresponding results from the inoculated network in the **main text** provided in parentheses for comparison.

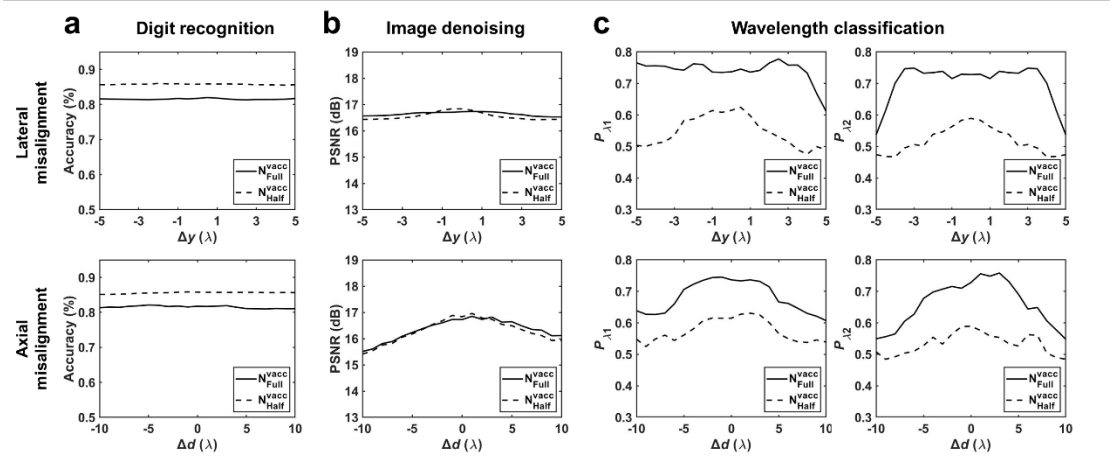

Fig. S4 Comparison of the performance of vaccinated networks with different error vaccination ranges. **a** Digit recognition tast. **b** Image denoising task. **c** Wavelength classification task.

As shown in **Fig. S4a**, for the digit recognition task in the **main text**, when the half-vaccinated networks under ideal condition, the classification accuracy are 85.78%(81.75%); Under lateral misalignments, the accuracy declines by an average of 0.10%(0.23%); Under axial misalignments, the accuracy declines by an average of 0.21%(0.34%).

As shown in **Fig. S4b**, for the image denoising task in the **main text**, when the half-vaccinated networks under ideal condition, the average PSNR are17.39dB(16.73dB); Under lateral misalignments, the average PSNR declines by an average of 0.29dB(0.09dB); Under axial misalignments, the average PSNR declines by an average of 0.59dB(0.43dB).

As shown in **Fig. S4c**, For the wavelength classification task in the **main text**, when the half-vaccinated networks under ideal condition, the P_{λ_1} are 61.49%(73.67%) and the P_{λ_2} are 58.94%(72.82%); Under lateral misalignments, the P_{λ_1} declines by an average of 7.14%(2.37%); the P_{λ_2} declines by an average of 7.47%(4.05%); Under axial misalignments, the P_{λ_1} declines by an average of 4.53%(5.90%); the P_{λ_2} declines by an average of 5.85%(8.02%);

The results indicate that, under the same training conditions, varying the range of training error vaccination affects both the ideal performance and error robustness of the vccinated network. Moreover, the observed effects differ across different task scenarios. Therefore, the inoculation range of training errors should be carefully determined based on the specific requirements and complexity of the task at hand.