Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Photonic machine learning with on-chip diffractive optics

**Tingzhao Fu**
Tsinghua University
**Yubin Zang**
Tsinghua University
**Yuyao Huang**
Tsinghua University
**Zhenmin Du**
Tsinghua University
**Honghao Huang**
Tsinghua University
**Chengyang Hu**
Tsinghua University
**Minghua Chen**
Tsinghua University
**Sigang Yang**
Tsinghua University https://orcid.org/0000-0002-2209-287X
**Hongwei Chen** ( chenhw@tsinghua.edu.cn )
Tsinghua University https://orcid.org/0000-0002-2952-2203

**Article**

**Keywords:**

**Posted Date:** May 13th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1550655/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

# Photonic machine learning with on-chip diffractive optics

**Tingzhao Fu, Yubin Zang, Yuyao Huang, Zhenmin Du, Honghao Huang, Chengyang Hu, Minghua Chen, Sigang Yang, Hongwei Chen***

*Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China*
**chenhw@tsinghua.edu.cn*

**Abstract:** Machine learning technologies have been extensively applied in high-performance information-processing fields. However, the computation rate of current hardware is severely circumscribed by the conventional Von Neumann architecture. Photonic approaches have demonstrated extraordinary potential for executing deep learning involving complex calculations. In this work, an on-chip diffractive optical neural network (DONN) based on a silicon-on-insulator (SOI) platform is proposed to perform machine learning tasks with high integration and low power consumption. To validate the proposed DONN, we fabricated 1-hidden-layer and 3-hidden-layer on-chip DONNs with footprints of 0.15 $mm^2$ and 0.3 $mm^2$ and experimentally verified their performance in a classification task on the Iris plants dataset, yielding accuracies of 86.7% and 90%, respectively. The proposed fully passive on-chip DONN provides a potential solution for accelerating future artificial intelligence hardware with enhanced performance.

Concomitant with the substantial progress made in semiconductor technologies and novel computing architectures [1-9], artificial neural network (ANN)-related machine learning applications are being extensively utilized in many fields, including computer vision [10], natural language processing [11], emotion detection [12], speech recognition [13], medical image analysis [14,15], and decision-making [16,17]. However, to solve complex tasks in a timely manner, ANNs require huge amounts of resources, both in terms of computing speed and energy consumption. In recent decades, optical neural networks (ONNs) have garnered tremendous interest, because of their advantages of low power consumption and ultrahigh computing bandwidth, which are unrivaled by their electronic counterparts [18-32]. Several implementations of ONNs have been proposed, including a coherent approach based on an integrated Mach-Zehnder interferometer (MZI) mesh [18,24,25,31], WDM processing with micro-ring modulators, and programmable routing enabled by a phase-change material (PCM) [20]. However, these architectures are burdened by their limited computational scales, which are significantly restricted by their footprint and energy consumption.

Recently, diffractive optical neural networks (DONNs) have been garnering much more attention for extending the optical computing capacity and lowering the power consumption by leveraging large-scale computations with the inherent parallel nature of optics [32-35]. This approach can map numerous neurons and connections onto optics, providing an even larger computational capacity than the conventional ONN architecture. However, mainstream DONNs are bulky because they are established on discrete diffractive components, causing significant difficulties integrating them into compact systems. In addition, complex calibrations between discrete devices may introduce additional errors.

In this work, we address the drawbacks of DONN by proposing an on-chip DONN architecture based on an integrated one-dimensional (1D) dielectric metasurface. The 1D dielectric metasurface consists of a series of silicon slots and represents the hidden layer (HL) in on-chip DONNs. To ensure that the pre-trained parameters can be mapped accurately onto physical structures, a silicon slot group is used as a single neuron [22]. To demonstrate the

capability of on-chip DONNs, we fabricated on-chip 1-hidden-layer DONN (DONN-1) and 3-hidden-layer DONN (DONN-3) based on the silicon-on-insulator (SOI) platform for the first time to the best of our knowledge. The spacing between the adjacent HLs was set as 250 μm, and the footprint of the on-chip DONN-1 and DONN-3 was 0.15 $mm^2$ and 0.3 $mm^2$, respectively. Then, we benchmarked their classification performance on the Iris plants dataset[36]. DONN-1 and DONN-3 yielded accuracies of 86.7% and 90%, respectively. We also propose an algorithm implemented by additional phase and power calibrations that compensates for fabrication errors, resulting in enhanced system noise resistance. The aforementioned method for designing and fabricating on-chip DONNs not only provides a solution for large-scale computation, but also overcomes the problem of complex alignment among the discrete components, which potentially paves the way for implementing future optical artificial intelligence accelerators and promotes the potential application of photonic integrated devices in many other aspects.

## Results

**On-chip DONN model.** The proposed on-chip DONN model consists of on-chip electromagnetic propagation, forward and error backward propagation, and a neuron-mapping process. The on-chip electromagnetic propagation model is modified based on the Huygens-Fresnel principle under restricted propagation conditions. It is an indispensable part of the on-chip DONN model, and can be described by Eq. (1):

$$w_i^m = \frac{1}{j\lambda}\cdot\left(\frac{1+\cos\theta_i}{2r_i}\right)\cdot\exp\left(j\frac{2\pi r_i n_{slab}}{\lambda}\right)\cdot\eta\exp\left(j\Delta\phi\right) \tag{1}$$

where $m$ represents the $m$-th layer of the network, $i$ represents the $i$-th neuron located at $(x_i, y_i)$ of layer $m$, $\lambda$ is the working wavelength, $j=\sqrt{-1}$ is an imaginary unit, $\cos\theta_i = (x-x_i)/r_i$, $r_i=\sqrt{(x-x_i)^2+(y-y_i)^2}$ is the distance between a neuron in layer $m-1$ and the $i$-th neuron in layer $m$, $n_{slab}$ is the effective refractive index (ERI) of the slab waveguide, $\eta$ is a specific coefficient of the amplitude and $\Delta\phi$ is a fixed phase delay[22]. The electric field evolution of the input signal propagation based on Eq. (1) is highly consistent with the simulation results of the 2.5D variational finite-difference time-domain (FDTD) solver (Supplementary Note 1.1).

Using an analytical expression of the on-chip electromagnetic propagation, the network structure parameters of the integrated DONNs can be pre-trained via forward and error backward propagation algorithms (Supplementary Note 1.2). Once the parameters of the on-chip DONNs are determined, these parameters can be mapped onto physical structures such as waveguides, grating couplers, multimode interferometer beam splitters, and silicon slots. Among the pre-trained parameters, the physical neuron-mapping process is the most critical. To ensure the reliability of the mapping process, the pre-trained phase value of a neuron is approximated by a silicon slot group composed of more than two identical slots. The length of the slots in each group can be calculated using Eq. (2):

$$L_{slot-i} = \frac{\Delta\varphi_i}{(n_{eff}-n_{slab})\cdot k_0} \tag{2}$$

where $L_{slot-i}$ is the length of the slots in the $i$-th group, $n_{eff}$ is the ERI of the slot group through which light passes, $n_{slab}$ is the ERI of the slab waveguide, $k_0=2\pi/\lambda$ is the wave number of light propagating in a vacuum, and $\Delta\varphi_i$ is the phase delay generated by the $i$-th slot group.

**DONN device architecture and design.** In on-chip DONNs, as depicted in Fig. 1 (a), the trainable parameters are the phase values, which need to be physically implemented by the diffractive units, where, each diffractive unit (DU) is a silicon slot group composed of three identical silicon slots, and we record this silicon slot group as a single neuron. For on-chip DONNs, the weight $w_{ij}$ connecting each neuron is fixed, and trainable phase values on distinct HLs are achieved by changing the size of the DUs.

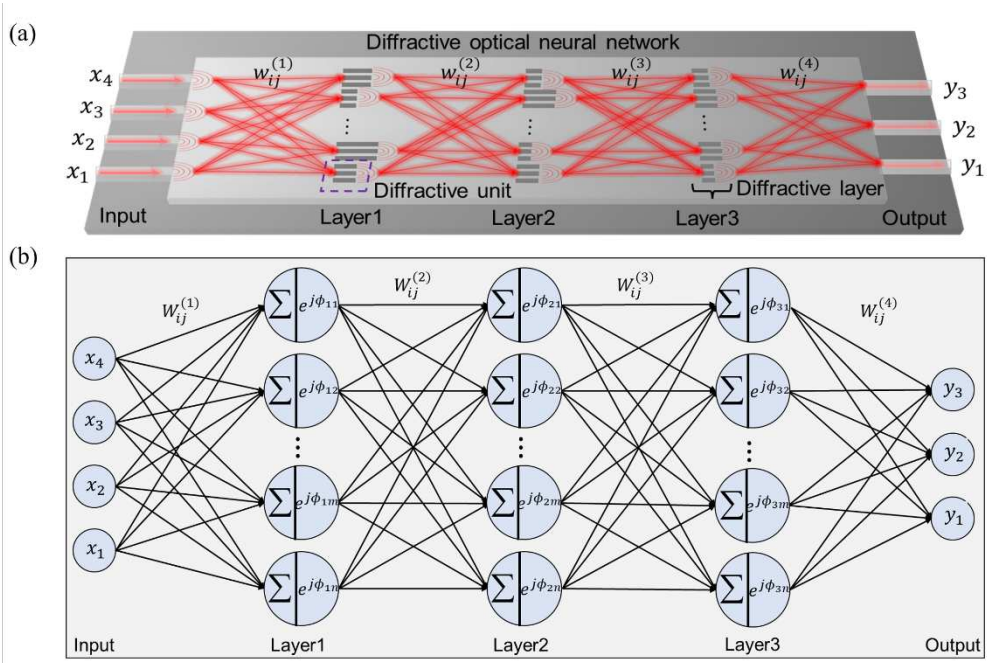


**Fig. 1.** (a) Schematic of an on-chip diffractive optical neural network (DONN), each diffractive unit on a given layer acts as a secondary source of a wave, the amplitude and phase of which are determined by the product of the input wave and the complex-valued transmission at that unit. Each diffractive unit is a silicon slot group composed of three identical silicon slots and represents a single neuron in the on-chip DONNs. (b) Logic diagram of (a), which mathematically describes the physical calculation process of the on-chip DONN.

The on-chip DONN-1 and DONN-3 were designed and verified via a classification task on the Iris plants dataset. First, the input features were modulated onto the phase of the input light, and then the dataset coded in the optical phase was used to train the parameters of the on-chip DONNs by adopting the adaptive moment estimation (Adam) optimizer. Then, the pre-trained parameters were mapped onto silicon-based structures (Supplementary Note 1.3). Additionally, to maximize the accuracy of the neuron-mapping process, the distance between the HLs was also considered [22].

In this work, the proposed on-chip DONN is all-optical and can be used to solve complex tasks through the interference of transmitted light. The working wavelength of the laser is 1.55 μm, and by fixing the width and thickness of the slots to 200 nm and 220 nm, free control of the phase delay caused by the slots can be achieved within the range from 0 to 2π by changing the length of the slots from 0 to 2.3 μm.

For the optimized on-chip DONNs, the length of the HLs was 280 μm along the Y-axis, and each HL contained 186 neurons (consisting of 558 rectangular slots). The distance between two successive HLs was 250 μm along the X-axis. The input signal is loaded onto the corresponding input waveguides and propagates 1010 μm through the inverse taper into the slab waveguide, and then propagates 250 μm through the slab waveguide to reach the first HL. After light exits the last HL, it also propagates 250 μm until it reaches the output layer of the

network, with three detector regions ("D1," "D2," and "D3") arranged in a linear configuration. Each detector is assigned a specific category. The width of each detector was 8 μm, and the distance between the centers of the two neighboring detectors was 70 μm.

A schematic of the on-chip DONN-3 is shown in Fig. 2. The on-chip DONN can implement interference and prediction mechanisms in a light-speed and passive manner, and it can be applied in many fields, including computer vision, natural language processing, and image recognition. A conceptual diagram of on-chip DONNs application scenario is shown in Fig. 3.

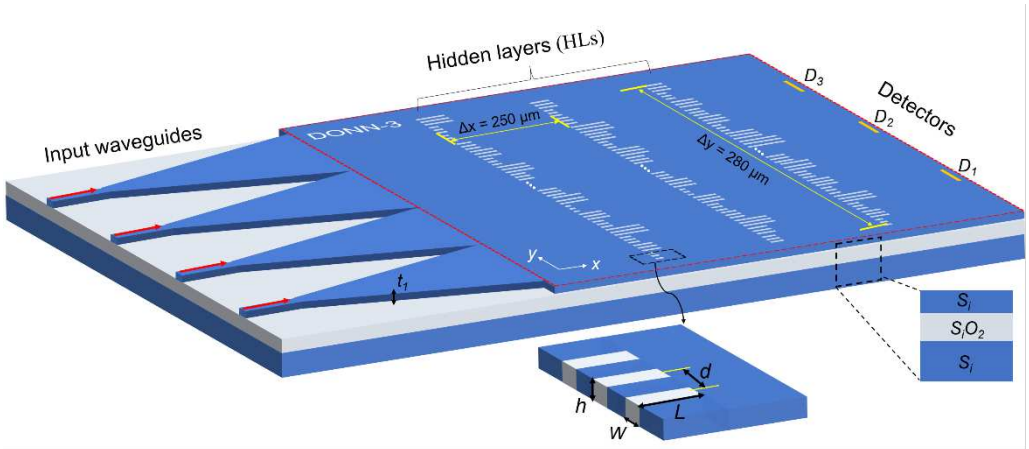


**Fig. 2.** Schematic of the on-chip DONN. It includes three hidden layers, and each neuron in the hidden layer consists of three identical slots representing a complex-valued transmission coefficient. The transmission coefficients of each layer can be trained by using deep learning to perform a function between the input and output plane of the network. Then, following fabrication of the on-chip DONN, it can perform the learned function at the speed of light in a passive manner. The center distance between the adjacent slots $d$ is 500 nm, the period of the silicon slot group is 1.5 μm, the width of the slot $w$ is 200 nm, the thickness of the slot $h$ is 220 nm, the length of the slot $L$ is determined by the pre-trained phase values based on Eq. (1), and the thickness of the light propagation layer $t_1$ is 220 nm; The distance between adjacent HLs is 250 μm along the X-axis, and the length of each HL is 280 μm along the Y-axis.

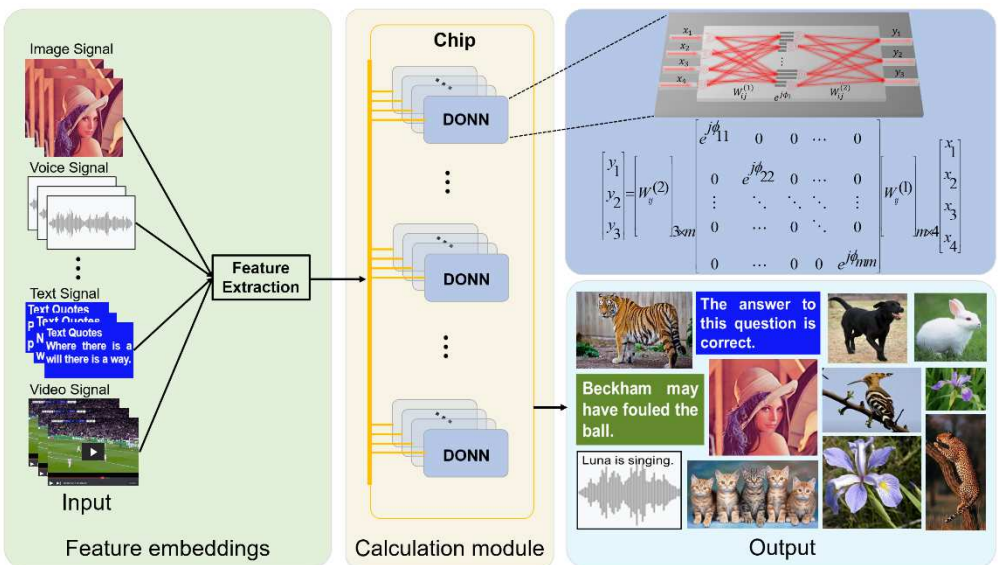


**Fig. 3.** Conceptual diagram of multi-channel on-chip DONN for various tasks. The features of different signals are extracted and encoded onto the phase, amplitude or polarization of light. Then, the input signals containing optical information are fed into the on-chip DONNs for subsequent calculations.

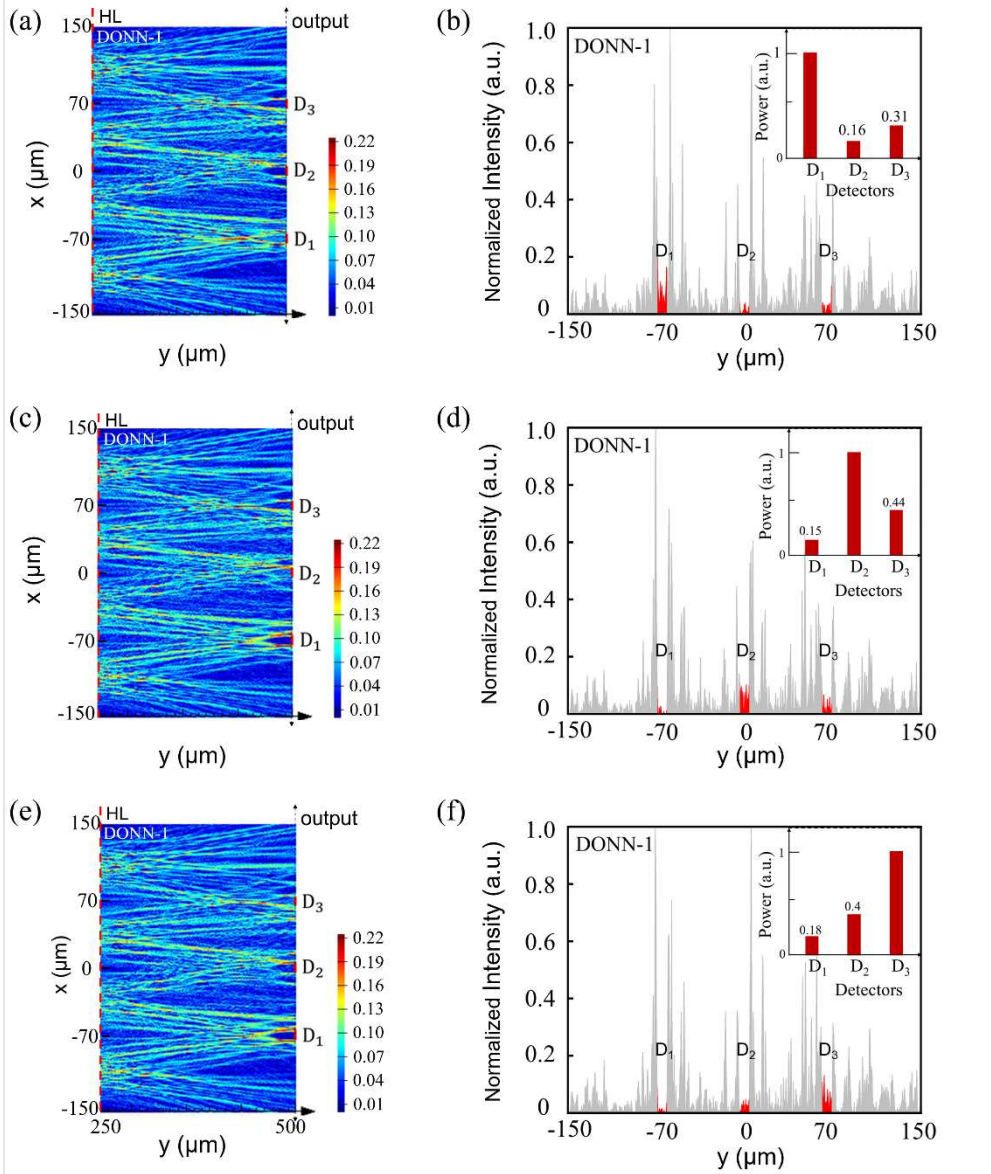


**Fig. 4.** Simulation results of the proposed on-chip DONN-1. In the picture, $D_1$, $D_2$, and $D_3$ represent the prediction of the different kinds of Irises; i.e., "Setosa," "Versicolor," and "Virginica," respectively. The red dotted line in the figure indicates the location of the hidden layer (HL). (a) Prediction of "Setosa", in other words, the power of $D_1$ should be greater than that of $D_2$ and $D_3$; here, the prediction result is correct. (b) Output waveform of (a); the red marks are the detection areas, with each detector 8 μm wide and the center spacing between adjacent detectors is 70 μm. (c) Prediction of "Versicolor"; the power of $D_2$ is greater than that of $D_1$ and $D_3$, thus the prediction result is correct. (d) Output waveform of (c). (e) Prediction of "Virginica"; the power of $D_3$ is greater than that of $D_1$ and $D_2$, thus the prediction result is correct. (f) Output waveform of (e).

**Numerical calculation and simulation.** Based on the on-chip DONN model, the on-chip DONN-1 and DONN-3 were optimized and utilized for classification on the Iris plants dataset. The dataset was divided into a training set and a testing set in a ratio of 8:2. The classification accuracies of on-chip DONN-1 and DONN-3 by numerical calculations were 86.7% and 90%, respectively. In addition, a 2.5D variational FDTD was used to verify the performance of on-chip DONN-1, and the classification accuracy of the simulation result was 86.7%. The matching score of the classification predictions between the 2.5D variational FDTD and the numerical calculation was 100%. Fig. 4 shows the simulation prediction process for the Iris species and the corresponding output waveforms of the FDTD simulation. This theoretical study also includes additional numerical calculation processes, relevant key parameters, and calculation results (Supplementary Note 2).

**Experiment.** As a proof of concept, the on-chip DONN-1 and DONN-3 were fabricated based on the SOI platform. The micrographs of the chips are shown in Figs. 5 (a) and (c). After processing and testing, the chips were packaged to facilitate subsequent experiments (Supplementary Note 4.1). Experimental tests were performed on this basis (Supplementary Note 4.2). A laser with a working wavelength of 1.55 μm was coupled into the waveguide via an input grating coupler, and then the input signal was loaded onto the phase of the light through four phase shifters. Finally, the modulated light interfered with the diffractive layers and was detected by the optical power meters at the output interface. The detected light was then transmitted to the CPU through analog-to-digital conversion, as shown in Fig. 5 (d). The numerical calculation and experimental testing results for on-chip DONN-1 and DONN-3 are listed in Table 1.

**Table 1 Numerical calculation and experimental testing results for on-chip DONNs**

| --- / Accuracy | Numerical calculation | Experimental test | |
|---|---|---|---|
| | | Uncompensated | Compensated |
| On-chip DONN-1 | 86.7% | 56.7% | 86.7% |
| On-chip DONN-3 | 90.0% | 60.0% | 90.0% |

The testing accuracies of on-chip DONN-1 and DONN-3 without compensation were 56.7% and 60.0%, respectively, which were significantly different from the theoretical calculations of 86.7% and 90%, respectively. Phase errors are generated during the fabrication process, and the error accumulation during light propagation significantly affects the performance of on-chip DONNs (Supplementary Note 3). To compensate for these errors, an algorithm compensation method, including phase compensation and power compensation, was exploited to reduce the negative impacts of machining errors (Supplementary Note 5 and Note 6). Meanwhile, the phase compensation stage was implemented based on the *in-situ* training procedure, during which a set of optimal voltage values can be obtained, and the output power was detected and recorded at this point. Here, the input signals were applied to the phases of light via the input voltages. After phase compensation, a traversal search method was adopted to find a set of optimal power compensation factors $(\alpha_1, \alpha_2, \alpha_3)$ to maximize the prediction accuracy of the dataset. Consequently, when external algorithm compensation was employed, the experimental testing accuracy of on-chip DONN-1 and DONN-3 was improved to 86.7% and 90%, respectively, as in the theoretical calculation. Fig. 6 shows the experimental testing results of on-chip DONN-1 and DONN-3 before and after the introduction of the error

compensation algorithm. From the compensated results, it can be observed that the compensation method is significantly effective, and the compensated results are consistent with the theoretical calculations.

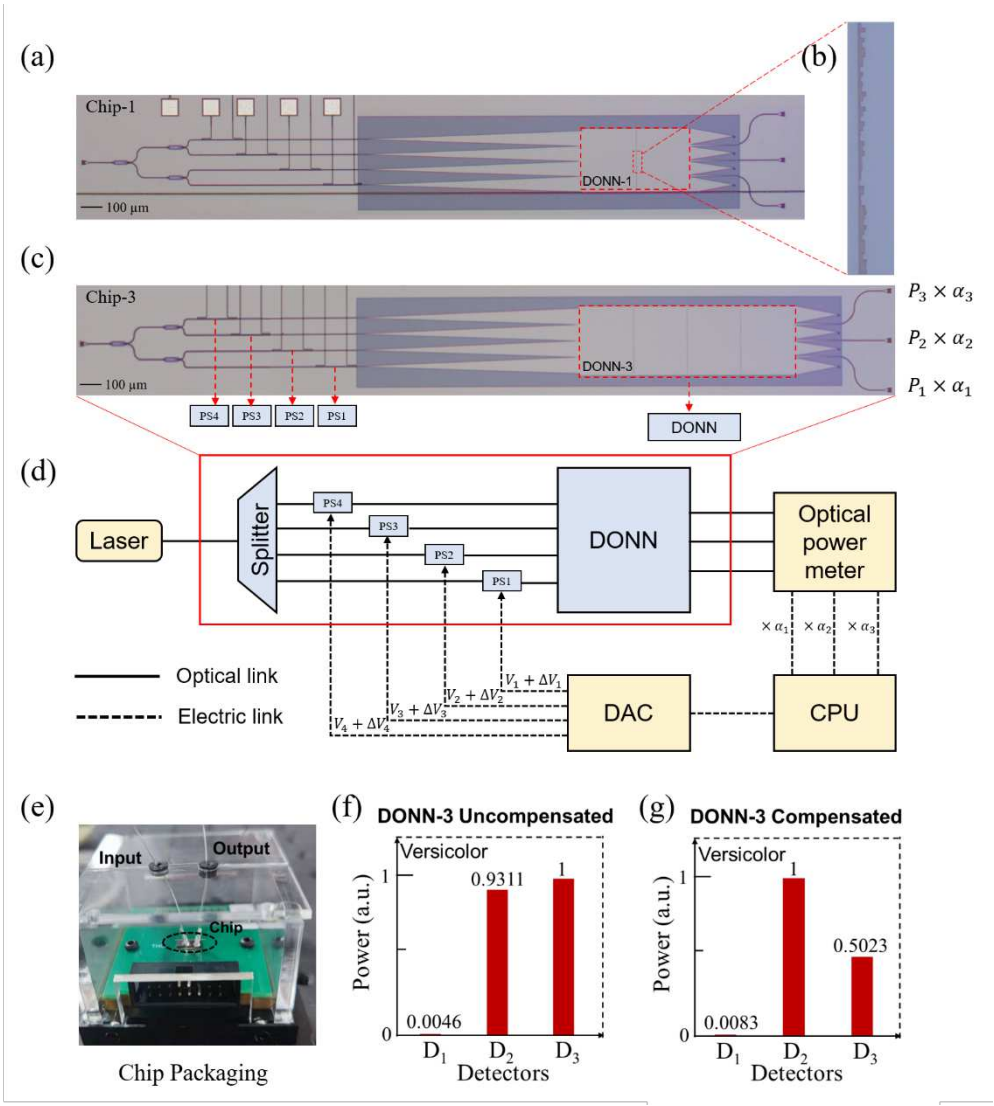


**Fig. 5.** (a, c) Micrographs of on-chip DONN-1 (a) and DONN-3 (c). (b) Close-up view of the partial slot array under the microscope's 100 × lens. (d) Experimental compensation testing process, where $V_i + \Delta V_i$ is the input voltage used to load the input signal and compensate for machining errors. $V_i$ is the loading voltage corresponding to the original signal of the dataset and $\Delta V_i$ is the compensation voltage found through the optimization algorithm in the phase compensation process. $\alpha_i$, which is a compensation factor for the output power to compensate for the fabrication errors. (e) Diagram of chip after packaging. (f, g) Experimental testing results for a sample in the testing set before (f) and after (g) compensation. The input features of the sample belong to "Versicolor," thus its correct prediction result is that the power of $D_2$ should be the largest. It can be seen from (f) and (g) that after compensation, the prediction result is corrected from the wrong result (f) to the correct result (g).

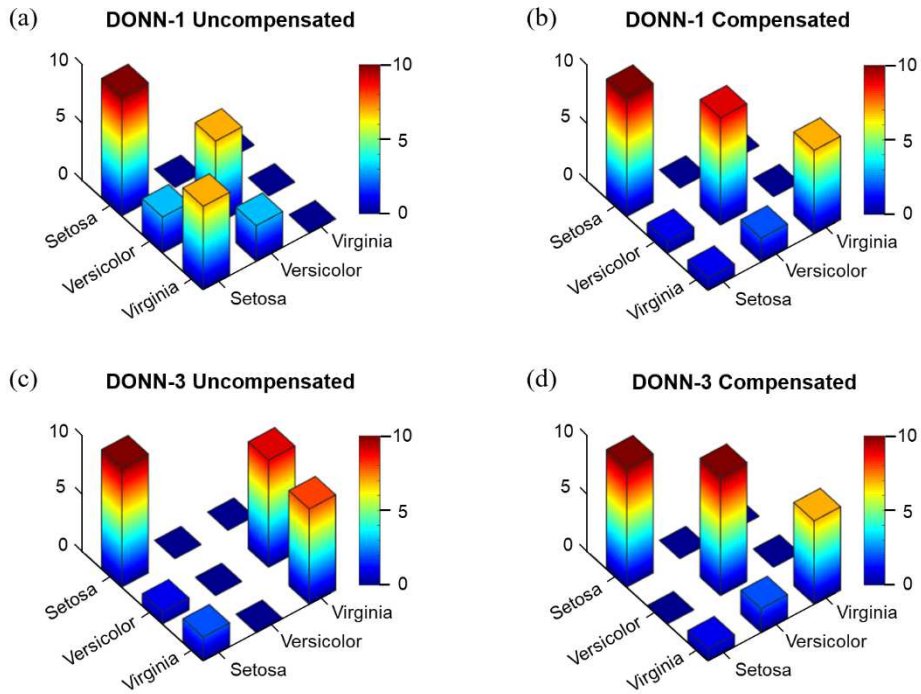


**Fig. 6.** Experimental testing results of on-chip DONN-1 and DONN-3. (a) Confusion matrix of the experimental results of DONN-1 without compensation. (b) Confusion matrix of the experimental results of DONN-1 after compensation. (c, d) Confusion matrixes of the experimental results of DONN-3 before (c) and after (d) the compensation.

## Discussion

**Experimental results analysis.** Table 1 indicates that the prediction accuracies of the on-chip DONN-1 and DONN-3 without algorithm compensation on the Iris plants dataset are 56.7% and 60.0%, respectively, which are quite different from the numerical calculation results of 86.7% and 90%, respectively. The difference between the experimental and numerical calculation results is attributable to the errors caused by the fabrication process. The working process of on-chip DONNs can be analytically expressed by Eq. (3) and Eq. (4):

$$Y_{cal} = D_{cal} X \tag{3}$$

$$Y_{chip} = D_{chip} X \tag{4}$$

where $Y_{cal}$ is the theoretical calculation result of the product of input $X$ and the transfer matrix $D_{cal}$, $D_{chip}$ is the transfer matrix of light propagating in the slab waveguide, and $Y_{chip}$ is the output electric field of the product of input $X$ and the transfer matrix $D_{chip}$. Because of inevitable machining errors, the error transfer matrix $D_{err}$ is bound to exist naturally after fabrication and, mathematically, $D_{err}$ is the difference between $D_{cal}$ and $D_{chip}$. Furtherly, $D_{err}$ results in the difference $P_{err}$ between the theoretically calculated power $P_{cal}$ and the detected power $P_{chip}$, that is, $P_{err} = P_{cal} - P_{chip}$. The external algorithm compensation aims to find a set of input voltage values and a power compensation factor $\alpha$ that minimizes the difference $P_{err}$, that is, minimizes the absolute value $|P_{cal} - \alpha \cdot P_{chip}|$; for example, the optimal solution $|P_{err}| = |P_{cal} - \alpha \cdot P_{chip}| \approx 0$. In this experiment, after compensation by the

algorithm, the prediction accuracies of on-chip DONN-1 and DONN-3 were improved to 86.7% and 90%, respectively. Both prediction accuracies after compensation by the algorithm are consistent with the theoretical calculation results.

**Effectiveness of pre-trained parameters.** To illustrate the effectiveness and importance of the pre-trained parameters of HL, 10 groups of HL parameters for different on-chip DONN-1 and DONN-3 were randomly generated, and the Iris plants dataset was used to test the performance of these on-chip DONNs. The prediction accuracy results are shown in Figs. 7 (a) and (b). It is evident that the prediction results of the on-chip DONNs with pre-trained HL parameters (serial number "1") are significantly higher than those of the on-chip DONNs with randomly generated HL parameters (serial numbers "2-11"). This proves that the pre-trained parameters are significantly effective and imperative.

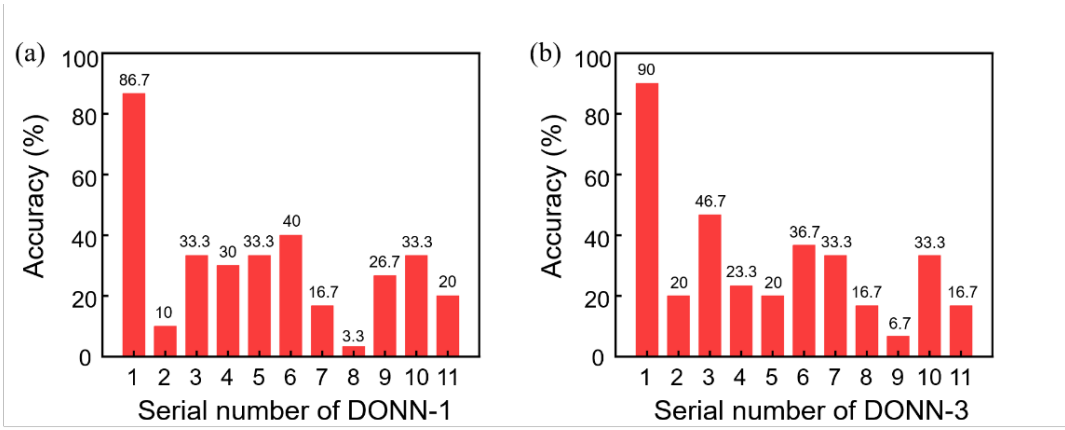


**Fig. 7.** Prediction accuracy results for (a) on-chip DONN-1 and (b) on-chip DONN-3. Serial number "1" indicates the result obtained on the premise of pre-trained hidden layer (HL) parameters. Serial numbers "2-11" indicate the results obtained on the premise of randomly generating HL parameters.

**Computation speed and energy efficiency.** When handling complex tasks, such as automatic driving and real-time missile tracking, ANNs with high speed and low energy consumption are necessary. Our on-chip DONN architecture takes advantage of processing big data at high speeds and low power consumption. Once all the parameters have been trained and mapped onto the physical structures, forward propagation computing is performed optically on a passive system. Assuming that our on-chip DONN has $m$ HLs and $N$ neurons at each HL, when it operates at a typical 100 GHz photodetection rate, the number of floating-point operations per second (FLOPS) to match the optical network is obtained using Eq. (5) [18]:

$$R = 2m \times N^2 \times 10^{11} \text{ FLOPS} \tag{5}$$

where $R$ is the number of operations per second, which is related to the number of HLs, number of neurons on each HL, and detection frequency of the photodetectors. Therefore, for the on-chip DONN-3, the computation speed is approximately $2.08 \times 10^{16}$ FLOPS calculated by Eq. (5), which is four orders of magnitude higher than the performance of modern GPUs, which typically perform at $10^{12}$ FLOPS [25]. Meanwhile, in the optical calculation process, the calculation delay was approximately 27.56ps (Supplementary Note 7.1). As regard energy consumption, our on-chip DONN only consumes energy during the signal loading and result detection stage, whereas the calculation process of the computing part is fully passive (Supplementary Note 7.2).

**Scalability and Reconfigurability.** Currently, the scalability of on-chip neural network is an obstacle. For example, interference ONNs based on MZIs [18] and pulse ONNs based on micro-ring resonators [20] cannot dramatically expand the number of neurons owing to the large footprint of each individual device. The on-chip DONN is a feasible method for solving this problem. In this work, we designed an on-chip DONN-3 with three HLs, and each HL included 186 neurons. According to the current design method, approximately 2000 neurons can be designed per square millimeter. Once the neuron mapping method is further improved, the number of integrated neurons can be significantly increased. For the reconfigurability of on-chip DONNs, PCM materials are candidate for future studies. For example, related work on PCM materials to realize reconfigurable networks has been reported [20,37].

**Performance of proposed DONN framework.** Table 2 compares the designed on-chip DONN-3 with other integrated ONNs. The matrix dimension is a key parameter for dealing with complex tasks; the size of the matrix dimension depends on the number of integrated neurons. For complicated tasks, the energy demand in the calculation process is greater. Therefore, a passive calculation process is imperative. From a comprehensive perspective, our proposed on-chip DONN architecture is a worthwhile choice. In Table 2, MR is the abbreviation for micro-ring and $N$ is the maximum number of neurons in the HL.

**Table 2 | Comparison between the on-chip DONN-3 and other integrated ONNs**

| Study | Node type | Size of HL ($N$) | Integration in theory (neurons/ mm$^2$) | Power consumption of computing part |
|---|---|---|---|---|
| Ref. [18] | MZI | 4 | < 10 | 1.92 W |
| Ref. [20] | MR covered with PCM | 4 | < 5 | Passive |
| Ref. [38] | MZI | 6 | < 10 | 3.92 W |
| Ref. [39] | MZI | 10 | < 20 | 7.7 mW |
| Ref.[40] | MMI | 5 | < 25 | 2.04W |
| Our work | Subwavelength unit | 186 | ~ 2000 | Passive |

**Conclusion.** A wholly passive fully optical on-chip DONN based on the SOI platform was proposed and fabricated for the first time. The on-chip DONN can perform complicated functions at a higher speed and with lower latency and power consumption than conventional ANNs. An inference task was used to demonstrate the performance of the on-chip DONN, and the results were excellent after introducing a compensation algorithm. Note that, nonlinear activations were not used in this study, and the results for inference tasks can be better if nonlinear activation functions are considered. Consequently, we will consider the implementation of nonlinear functions on-chip in combination with PCM in future works. Furthermore, compared with other ONNs, the proposed on-chip DONN has the advantages of a simple structure design, all-optical passive operation, and massive-scale neuron integration. This on-chip DONN architecture is a potential solution for accelerating future artificial intelligence hardware with enhanced performance.

## Methods

**Device fabrication.** The entire on-chip DONN was fabricated on a SOI (100 substrate) platform with a 220 nm thick silicon (Si) top layer and a 3 μm thick buried oxide. The slots were created by etching the 220 nm Si film layer, and then a 2 μm thick silicon dioxide ($SiO_2$) protection layer was deposited on the device layer. A thin layer of titanium nitride (TiN) was deposited as a resistive layer for the heaters. A metal film of AlCu (Cu:0.5%) was patterned as the electrical connection to the electrodes and heaters. The fabrication was performed at the Institute of Microelectronics, Chinese Academy of Sciences.

**Optical measurements.** A continuous-wave tunable semiconductor laser with a polarization controller was used to launch light onto the chip (10 mW). The fiber-grating coupler loss was optimized to 5 dB per input/output facet. The output was monitored using two dual-channel optical power meters, and the minimum power detection limit was -75dB. An external auxiliary circuit was provided by a DC dual-tracking voltage-stabilizing source (DH1718E-5, 0-35V).

**Numerical simulations.** The training process of the Iris flower was conducted in Pytorch, which is a package for python. The light diffraction connection in the process of forward and error backward propagation followed the modified Huygens-Fresnel principle. The data in the Iris plants dataset were encoded into the light phase, ranging from 0 to $2\pi$. A 2.5-dimensional variational FDTD method (http://www.lumerical.com/tcad-products/fdtd/) was used to simulate the optical field distribution and on-chip DONN systems. A conformal mesh with a spatial resolution of less than 1/10 of the smallest feature size was applied.

## Data availability

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

## References


1. Liu, Y.Q. *et al.* Effective scaling of blockchain beyond consensus innovations and Moore's law: Challenges and opportunities. *IEEE Systems Journal* (2021).
2. Taylor, M.B. The evolution of bitcoin hardware. *Computer* **50**, 58-66 (2017).
3. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-+ (2016).
4. Chen, Y.H. *et al.* Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits* **52**, 127-138 (2017).
5. Misra, J. *et al.* Artificial neural networks in hardware: A survey of two decades of progress. *Neurocomputing* **74**, 239-255 (2010).
6. Esser, S.K. *et al.* Convolutional networks for fast, energy-efficient neuromorphic computing. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 11441-11446 (2016).
7. Poon, C.S. *et al.* Neuromorphic silicon neurons and large-scale neural networks: Challenges and opportunities. *Front Neurosci* **5**, 108 (2011).
8. Graves, A. *et al.* Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471-+ (2016).
9. Shafiee, A. *et al.* ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, 14-26 (2016).
10. Krizhevsky, A. *et al.* ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60**, 84-90 (2017).
11. Ananthanarayana, T. *et al.* Deep learning methods for sign language translation. *ACM Transactions on Accessible Computing* **14**(2021).
12. Byun, S.W. *et al.* Emotion recognition from speech using deep recurrent neural networks with acoustic features. *Basic & Clinical Pharmacology & Toxicology* **123**, 43-44 (2018).
13. Graves, A. *et al.* Speech Recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6645-6649 (2013).
14. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60-88 (2017).

15. Apostolidis, K.D. *et al.* A survey on adversarial deep learning robustness in medical image analysis. *Electronics* **10**(2021).
16. Kruglov, I. *et al.* Quantile based decision making rule of the neural networks committee for ill-posed approximation problems. *Neurocomputing* **96**, 74-82 (2012).
17. Ozkan, G. *et al.* Comparison of neural network application for fuzzy and ANFIS approaches for multi-criteria decision making problems. *Applied Soft Computing* **24**, 232-238 (2014).
18. Shen, Y.C. *et al.* Deep learning with coherent nanophotonic circuits. *Nature Photonics* **11**, 441-+ (2017).
19. Lin, X. *et al.* All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004-+ (2018).
20. Feldmann, J. *et al.* All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208-+ (2019).
21. Zarei, S. *et al.* Integrated photonic neural network based on silicon metalines. *Optics Express* **28**, 36668-36684 (2020).
22. Fu, T.Z. *et al.* On-chip photonic diffractive optical neural network based on a spatial domain electromagnetic propagation model. *Optics Express* **29**, 31924-31940 (2021).
23. Feldmann, J. *et al.* Parallel convolutional processing using an integrated photonic tensor core (vol 589, pg 52, 2021). *Nature* **591**, E13-E13 (2021).
24. Fang, M.Y.S. *et al.* Design of optical neural networks with component imprecisions. *Optics Express* **27**, 14009-14029 (2019).
25. Williamson, I.A.D. *et al.* Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *IEEE Journal of Selected Topics in Quantum Electronics* **26**(2020).
26. Miscuglio, M. *et al.* All-optical nonlinear activation function for photonic neural networks. *Optical Materials Express* **8**, 3851-3863 (2018).
27. Zuo, Y. *et al.* All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132-1137 (2019).
28. Bueno, J. *et al.* Reinforcement learning in a large-scale photonic recurrent neural network. *Optica* **5**, 756-760 (2018).
29. Khoram, E. *et al.* Nanophotonic media for artificial neural inference. *Photonics Research* **7**, 823-827 (2019).
30. Mengu, D. *et al.* Analysis of diffractive optical neural networks and their integration with electronic neural networks. *IEEE Journal of Selected Topics in Quantum Electronics* **26**(2020).
31. Hughes, T.W. *et al*. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica* **5**, 864-871 (2018).
32. Yan, T. *et al.* Fourier-space diffractive deep neural network. *Physical Review Letters* **123**(2019).
33. Miscuglio, M. *et al.* Massively parallel amplitude-only Fourier neural network. *Optica* **7**, 1812-1819 (2020).
34. Wu, Z.C. *et al*. Neuromorphic metasurface. *Photonics Research* **8**, 46-50 (2020).
35. Zhou, T.K. *et al.* Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics* **15**, 367-373 (2021).
36. Blake, C.L. & Merz, C.J. UCI repository of machine learning databases, 1998. (1998).
37. Delaney, M. *et al.* Nonvolatile programmable silicon photonics using an ultralow-loss Sb2Se3 phase change material. *Science Advances* **7**, eabg3500 (2021).
38. Zhang, H. *et al.* An optical neural chip for implementing complex-valued neural network. *Nature Communications* **12**, 1-11 (2021).
39. Zhu, H.H. *et al.* Space-efficient optical computing with an integrated chip diffractive neural network. *Nature Communications* **13**, 1044 (2022).
40. Zhao, X. *et al.* On-chip Reconfigurable Optical Neural Networks. (2021).

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) (62135009) and the National Key Research and Development Program of China (2019YFB1803500).

## Author contributions

T.F. and H.C. conceived the idea. T.F. developed the design principle and numerical simulations. T.F., Y.Z. and Z.D. did the experiments. T.F., Y.Z., H.H., Y.H., Z.D., C.H., S.Y., M.C. and H.C. involved in the discussion, theoretical analysis and data analysis. T.F. prepared the manuscript. Y.H., Y.Z., and H.C. revised the manuscript. H.C. supervised and coordinated all the work. All authors commented on the manuscript.

## Competing interests

The authors declare no competing interests.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryPhotonicmachinelearningwithonchipdiffractiveoptics.r.docx