

Supplementary information for

Ultra-compact multi-task processor based on in-memory optical computing

Wencan Liu^{1,2}, Yuyao Huang^{1,2}, Run Sun^{1,2}, Tingzhao Fu³,
Sigang Yang^{1,2}, and Hongwei Chen^{1,2,*}

¹Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

²Beijing National Research Center for Information Science and Technology (BNRist)

³Hunan Provincial Key Laboratory of Novel Nano Optoelectronic Information Materials and Devices,
College of Advanced Interdisciplinary Studies, National University of Defense Technology,
Changsha 410073, China

*Correspondence: chenhw@tsinghua.edu.cn

1 Supplementary note 1: Deep regression neural network setup

In our previous studies [1–3], the process of neuron mapping was typically reduced to the phase alteration of the optical field produced by the length variations of silica slots. However, the relationship between the optical field and these slots is governed by a complex function that is challenging to articulate in a direct mathematical expression, therefore we use slots group to represent one single neuron. We also increase the layer spacing to mitigate the impact of the angle of incident light, thereby stabilizing the effective refractive index (ERI) and facilitating a more precise mathematical analysis. In the present research, we employ the DRNN empowered by ComplexPyTorch [4–6] to approximate the intricate function representing light propagation within a line of slots, referred as metaline [7]. Assuming that the width, thickness and lattice constant of the slots are fixed, and their length is arranged as L_{slot} . The maximum length of each slot is L . The starting position of a metaline is x_0 , and the electrical field distribution of the optical field at x_0 is E_{x_0} , so the field distribution of the optical field at $(x_0 + L)$ μm can be calculated according to Eq. (S1):

$$E_{x_0+L} = F(E_{x_0}, L_{slot}) \quad (1)$$

where F is the complex function of light propagating in the metaline, and when the width, thickness and lattice constant of the slot are fixed, the optical field succeeding the metaline $E_{x_0} + L$ is only related to the optical field preceding the metaline and the length of the slots. Therefore, it is only necessary to find the mapping function F , giving the length of the slots in the metaline and the optical field preceding the metaline, then the optical field succeeding the metaline can be quickly calculated. With this method, the neuron mapping process can be represented by neural network, which no longer requires slot group and can take effect at any layer spacing. In pursuit of the mapping function F , the approach involves using the deep regression neural network (DRNN) to approximate the light propagation process within the metaline. For this work, we employ 2.5D FDTD to generate both training and testing datasets. We utilize the ComplexPyTorch library [41,42] within the Pytorch platform to build the model's architecture due to its capability to handle complex-valued data. Assuming that the data size is N , in the training process, complex-valued mean square error (CMSE) is used as the loss function, which is defined as Eq. (S2):

$$CMSE(x, y) = \frac{1}{N} \sum_N [(x_{real} - y_{real})^2 + (x_{imag} - y_{imag})^2] \quad (2)$$

where x is the output of DMRM and y is the ground truth of var-FDTD simulation. Since x and y are both complex-valued, x_{real} and x_{imag} are the real and imaginary part of x , and the same for y_{real} and y_{imag} . Goodness of fit (R^2) is chosen as the measure of fitting effect of the DMRM, which can be calculated as Eq. (S3):

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{1}{N} \sum_N \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (3)$$

where y_i is the ground truth, $\hat{y}_i$ is the DMRM output and $\bar{y}_i$ is the Average value of y_i . Here consider both the real part and the imaginary part of the predicted optical field and ground truth.

With this method, to validate the effectiveness of the trained DRNN, the simulational results for w/ diffraction units and w/o diffraction units are shown as Fig. S1. In the training process, the loss curves of training set and verification set are shown in Fig. S2(a). The goodness of fit (R^2) curves of the validation

set is shown in Fig. S2(b). Fig. S2(c) illustrates that the post-training DRNN prediction results closely align with the simulation results. This high degree of similarity, especially evident in the overall trend, peaks, and valleys of the waveform, provides substantial evidence of the effectiveness of the DRNN.

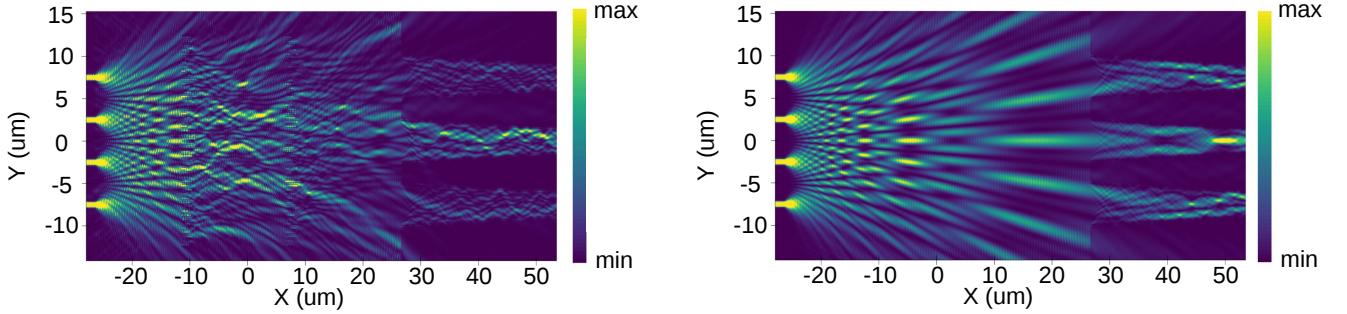


Figure S1: Simulation propagation results for w/ diffraction units (left) and w/o diffraction units (right), highlight the modulation ability of the SOC in an ultra-compact structure.

2 Supplementary note 2: Insertion loss and experimental errors

In our experiment, the operating wavelength of the incident laser was 1550 nm, and the input power was measured to be 5 dBm. The total power detected at the three output ports of SOC was approximately -16.5 dBm, indicating an overall insertion loss of approximately 21.5 dB. In order to measure the loss of the designed SOC chip, we divide the insertion loss into two parts: coupling loss between fibers and the chip and waveguide loss, and the loss of the diffraction core of SOC chip. For this purpose, we designed a contrast structure on the chip, including the same coupling structure and waveguide design, and the length of the waveguide is also approximately the same. After measurement, the loss of this control structure is approximately 13 dB, which means the total loss is 13 dB when the incident light does not pass through the diffraction core of SOC chip. Combining the overall insertion loss of 22.5 dB, it can be concluded that the insertion loss in the diffraction core is around 8.5 dB. In our previous works, due to the utilization of approximate modeling methods, the overall size of the chip is relatively large, and the measured insertion loss in the diffraction core reached more than 20 dB. Compared to our previous method, the proposed modeling method not only improves the chip integration, but also reduces the loss of diffraction core by more than 12 dB. Thus, smaller size of the chip means higher detection power values, thereby reducing the impact of noise, which also demonstrates the superiority of high integration chips and the proposed method. The comparison of insertion loss of the proposed SOC and previous on-chip DONN is shown in Table. S1.

We also found that there are errors in insertion loss and experimental results between the fabricated SOC chip and the simulation design. We believe that the errors may come from the signal loading, chip fabrication, and signal detection stages. By utilizing more advanced processing equipment, the errors introduced during chip fabrication can be fundamentally reduced. Additionally, incorporating perturbations into the data input during network training can further enhance the network's robustness. Finally, more advanced network training methods will also improve the error tolerance of the SOC chips.

Table S1: **Insertion loss of the proposed SOC and previous on-chip DONN.**

diffraction-based chip	distance between adjacent layers	Insertion loss
DONN [1]	300 um	33.2 dB
OCU [8]	300 um	21.4 dB
C-DONN [7]	15 um	6.9 dB (simulation)
SOC	15 um	8.5 dB

63 3 Supplementary note 3: Computational energy efficiency

64 The power consumption of SOC chip mainly comes from five parts: external laser source, on-chip thermal
65 phase shifters, external current source, photodetectors and digital backend. We conclude the power consumption of SOC chip and LTNs in Table. S2 as listed in the follow:

Table S2: **Estimated power consumption of SOC + LTN system.**

Components	Power (W)
External laser source (benchtop)	100
External current source (benchtop)	120
Digital controller and LTNs (powered by intel core i7-11800H)	50
On-chip thermal phase shifters	0.01
Photodetectors	1.5
Total power consumption	271.51

66 Note that the average resistance of the utilized TiN heaters was measured as 281.4Ω . Consequently, for
67 the on-chip thermal phase shifters, each with an average tuning current of 3 mA, were employed, generating
68 a power consumption of 0.01 W. Furthermore, 3 photodetectors, each with a power consumption of 0.5
69 W, were utilized, resulting in an overall power consumption of around 1.5 W. The digital processor was
70 responsible for feature mapping and launching LTNs, with an estimated power consumption of 30 W,
71 using an Intel core i7-11800H CPU. Consequently, the total power, including benchtop instruments such
72 as an external laser source with a power consumption of 100 W and an external current source with a
73 power consumption of 120 W, can be calculated as 271.51 W. As the calculated computational throughput
74 of the SOC at 10 Gbaud modulation speed is 304 TOPS and $191 \text{ POPS} \cdot \text{mm}^{-2}$, The energy efficiency
75 of the SOC chip can then be calculated as $1.12 \text{ TOPS} \cdot \text{W}^{-1}$ and $703.5 \text{ TOPS} \cdot \text{W}^{-1} \cdot \text{mm}^{-2}$. The energy
76 efficiency of SOC is not high due to the peripheral power consumption, which will not obviously increase
77 as the scale of SOC increases.

79 The in-memory computation proposed involves integrating the computing weights directly into the
80 chip using silica slots, allowing them to interact directly with the input light when data is loaded. The
81 computing process of SOC chips only requires loading the data without loading the weights that need to be
82 multiplied or added. Therefore, once the signal is loaded during the computing process (phase modulation
83 is used in this work), the subsequent computing process is power-consumption free, which is significant.
84 The digital parameters of SOC+LTN network and electronic network in multi-dataset classification task is
85 6 and 59, respectively. The digital parameters of SOC+LTN network and electronic network in regression
86 task is 18 and 36, respectively. Therefore, the in-memory computing architecture reduces the power-
87 intensive digital computation in classification and regression tasks of this work by around 90% and 50%,
88 respectively. Although the energy consumption of signal loading and external devices supporting the
89 signal loading process in this work is relatively high, we believe that the energy consumption can be
90 solved through better fabrication methods and more specialized and lightweight external loading devices.
91 The passivity of SOC chip computing process is significant.

92 **4 Supplementary note 4: Parameter settings of SOC**

93 The proposed SOC chips with 4 inputs and 3 outputs and two layers each containing 50 subwavelength
94 silica slots are designed for the following reasons: First, the inputs and outputs are designed aiming to
95 certain tasks. The 4 inputs and 3 outputs correspond to three multi-kernel convolutions with a 2×2 kernel
96 size for the regression task discussed in regression tasks. The input and output dimensions are adequate
97 for the multi-task scenarios examined in classification tasks. Second, the 2 hidden layers, each containing
98 50 subwavelength silica slots, were optimized for the task. Our analysis indicated that increasing the
99 number of layers and neurons does not significantly enhance accuracy, as validated in Fig. S10.

100 **5 Supplementary note 5: Scalability of SOC**

101 In this work, we have significantly enhanced the scalability of on-chip DONN, innovatively enabling the
102 same chip to adapt to different tasks. In this section, we will discuss the scalability challenges that SOC
103 chips encounter when addressing more complex tasks.

104 The limitation of input/output dimensions has been a persistent challenge for on-chip optical comput-
105 ing systems. While it is relatively straightforward to expand the input-output dimensions and the number
106 of slots, hidden layers within a certain range [1], scaling to higher dimensions poses significant challenges,
107 particularly regarding size and insertion loss. However, several solutions to improve the scalability of SOC
108 are available, such as parallel multi-chip processing and the use of one-hot encoding for limited outputs,
109 which can map to a larger variety of output types [9] (e.g., 8-channel one-hot encoding can generate 256
110 output types). Although these methods may lead to a slight reduction in accuracy, they significantly
111 increase the number of output channels and are applicable to our system.

112 **6 Supplementary note 6: Operating wavelength of SOC**

113 Our chip design approach involves first using computer simulations, tailored to the intended application
114 scenario, to determine the specific parameters of its computational units. Then we proceed with fab-

115 rication, producing a chip capable of delivering the required computational functionality. Because the
 116 operational wavelength influences the computational function of the chip’s structural units, it is essential
 117 to select the appropriate operating wavelength during the simulation stage. This chosen wavelength is
 118 subsequently applied in the experimental phase.

119 To ensure that our network training and structural design processes are effective, we first established
 120 the operating wavelength before fabrication. Based on this predetermined wavelength, we utilized Eq. (3)
 121 and (4) in the manuscript to simulate the on-chip optical propagation and computation. In this study,
 122 we selected 1550 nm, a standard communication wavelength, as our system’s operating wavelength from
 123 the outset of our design and simulation. So, in the experimental process after chip fabrication, we choose
 124 the 1550 nm wavelength that matches the design process.

125 After fabrication, we also experimentally performed a wavelength-scan of the two SOC chips, the
 126 results of which are presented in Fig. S11, which indicates that at the initially chosen wavelength of
 127 1550 nm, the system achieved performance consistent with simulation while maintain lower loss. Sig-
 128 nificant deviations in wavelength can lead to substantial discrepancies between the chip’s computational
 129 performance and its intended functionality, while also causing a marked increase in overall insertion loss.

130 7 Supplementary note 7: Fabrication process of SOC

131 The SOC chip was fabricated on an SOI platform with a 220nm top layer silicon (Si) and a 2 μ m dioxide
 132 layer via the process of EBL. The subwavelength metalines were fabricated by etching 220nm Si film and
 133 a 2 μ m thick silica upper cladding was deposited on the Si film. In the optimized configuration, each
 134 metaline, measuring 30 μ m in width (W), is perceived as a hidden layer and distributed along the Y-axis.
 135 As depicted in Fig. S12, the length and width of a metaline are set at 3 μ m and 30 μ m respectively.
 136 Each slot is characterized by a thickness (n) of 220 nm and a width (m) of 200 nm. The slots are spaced
 137 300 nm apart (q), and the maximum length of each slot (p) is equal to the metaline’s width (i.e., 3 μ m).
 138 Fig. S12 (a) shows the detailed fabrication parameters and the cross sectional images of the SOC chip we
 139 fabricated. Fig. S12 (b) shows some key parameters of our designed diffractive core of SOC. The length
 140 of the slab (Si) waveguide is 53 μ m and the width is 30 μ m. The thickness of each slot is $n = 220$ nm,
 141 the width is $m = 200$ nm, the max length is $p = 3$ μ m, and the distance between adjacent slots is $q =$
 142 300 nm.

143 8 Supplementary note 8: Stability and fabrication errors of SOC

144 Regarding the stability and errors of the chip and overall system during the experiment, we categorize
 145 the errors into two types: instability-related errors during experimental system testing and system errors
 146 introduced by chip fabrication and manufacturing processes. Concerning the stability of the experimental
 147 system, our testing system demonstrates relatively high stability, and thanks to the structural design, our
 148 thermal modulation operates reliably without the need for additional TEC device. To verify the reliability
 149 of the SOC chip and the experimental system, we measured the output results of the SOC chip with the
 150 same set of data under two different time with reconnection of the experimental system and compared
 151 the two, as shown in Fig. S13. The results indicate that the system maintains good stability and is not
 152 affected by fluctuations of the testing environment. Concerning the fabrication error of the SOC chip, we

153 acknowledge that fabrication errors like the EBL patterned lines may lead to slight structural variations
 154 in different chips. However, during the simulation process, our parameter optimization algorithm is based
 155 on gradient descent, which endows the optimized model with a certain degree of generalization ability
 156 and robustness against noise. Additionally, our team has conducted multiple fabrications of on-chip
 157 diffractive optical neural networks within the same facility, and through several iterations of technology
 158 and accumulated manufacturing experience, we have minimized errors induced by the fabrication process.
 159 During the experiments, we observed that the impact of inter-chip variations on the final results was not
 160 significant.

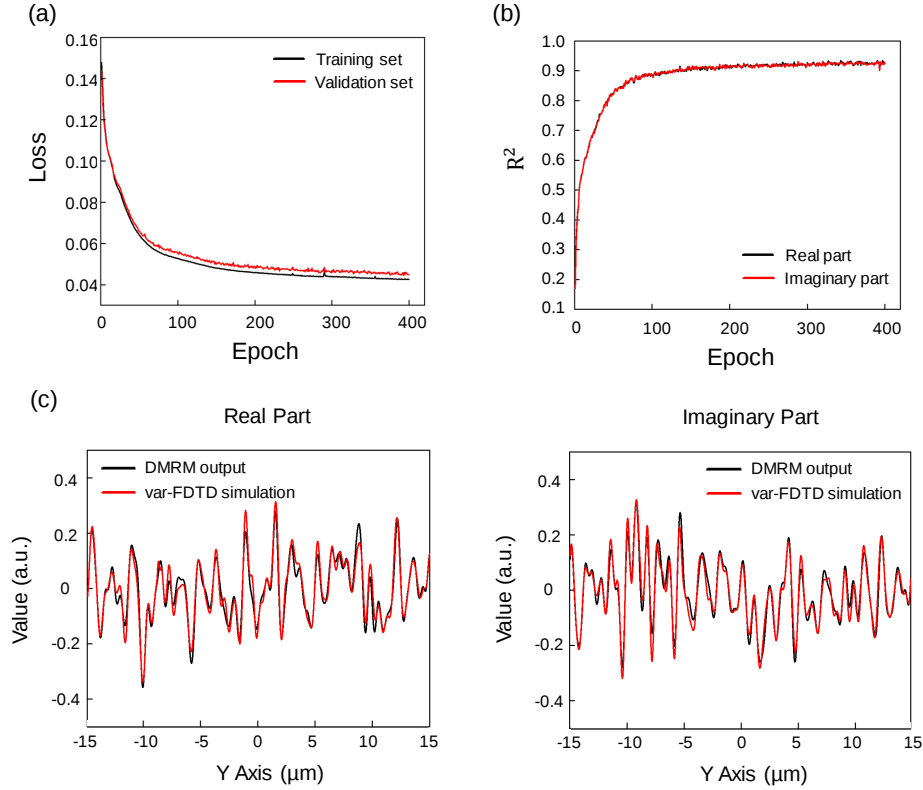


Figure S2: (a) The loss curves on the training set (black line) and validation set (red line) for the DMRM during the fitting procedure. (b) The goodness of fit (R^2) curves on the real part (black line) and imaginary part (red line) of the validation set during the learning procedure. (c) The comparison between the DRNN output and var-FDTD simulation results after the training process for the same sample in testing set.

Table S3: comparison between four network structures

structure	test accuracy of seed dataset	test accuracy of penguins dataset	test accuracy of iris dataset	number of digital parameters
SOC w/ aligning	75.9%	62.3%	83.3%	0
Light electrical network	82.5%	95.6%	93.3%	11
electrical network	92.5%	98.6%	96.7%	59
SOC + LTN	92.5%	98.6%	96.7%	6

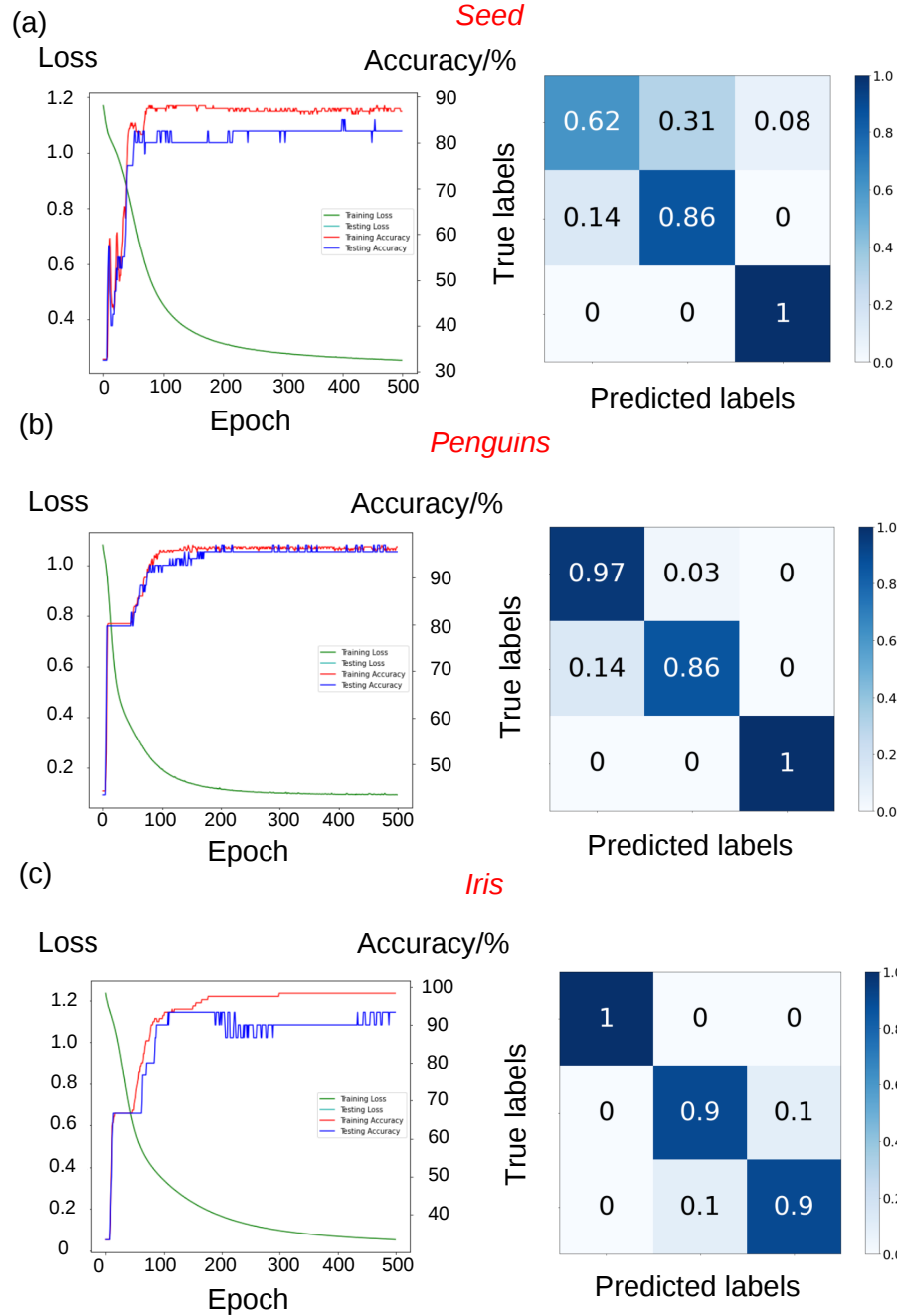
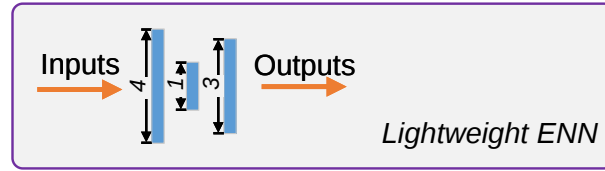


Figure S3: The accuracy and loss curves and confusion matrices of testing dataset for Light electrical net structure in (a) seed, (b) penguins, (c) iris dataset, respectively. The test accuracies for the 3 datasets are 82.5%, 95.6% and 93.3%, respectively.

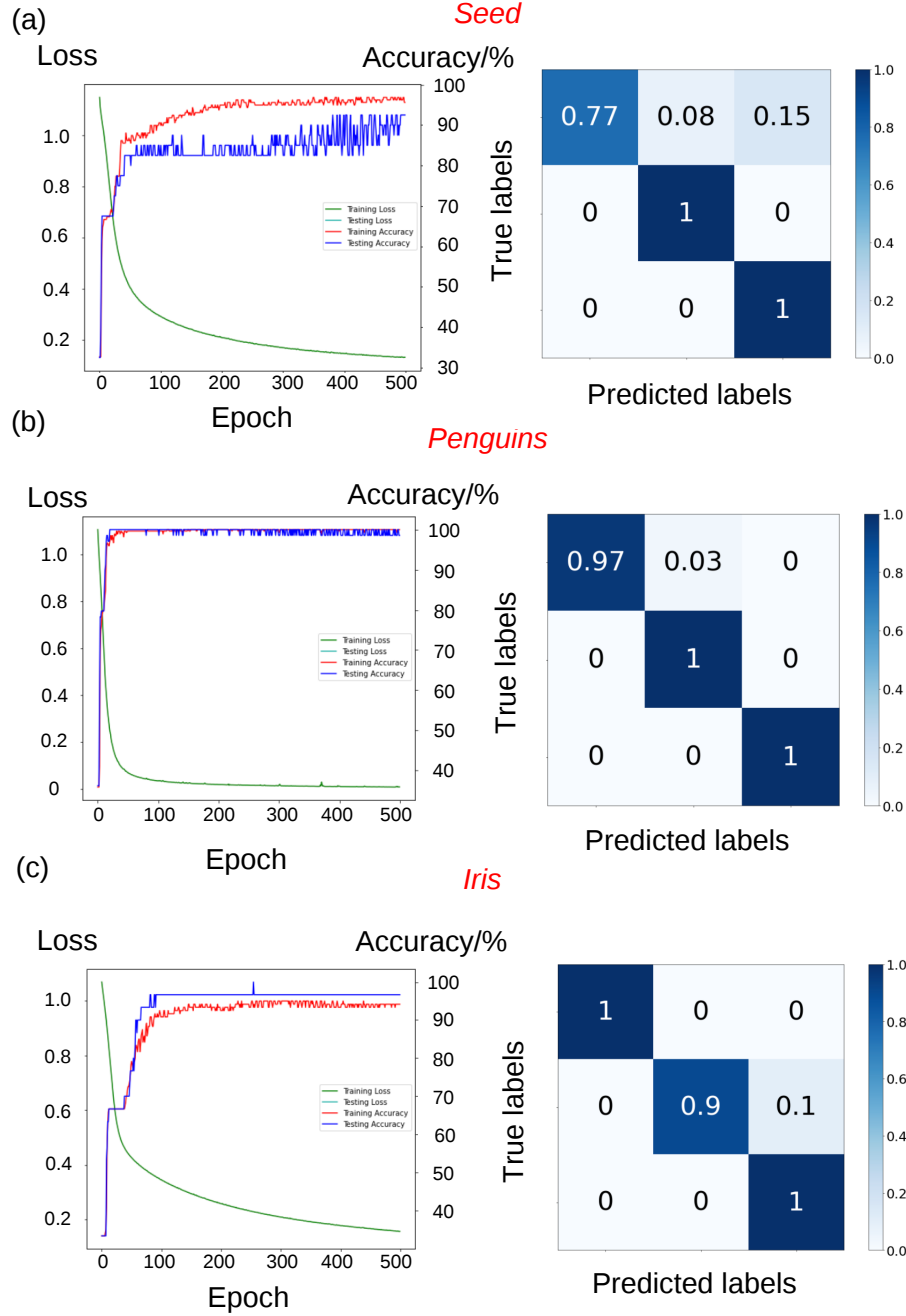
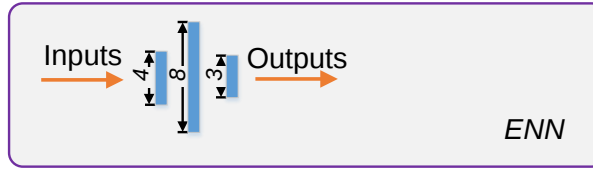


Figure S4: The accuracy and loss curves and confusion matrices of testing dataset for electrical net structure in (a) seed, (b) penguins, (c) iris dataset, respectively. The test accuracies for the 3 datasets are 92.5%, 98.6% and 96.7%, respectively.

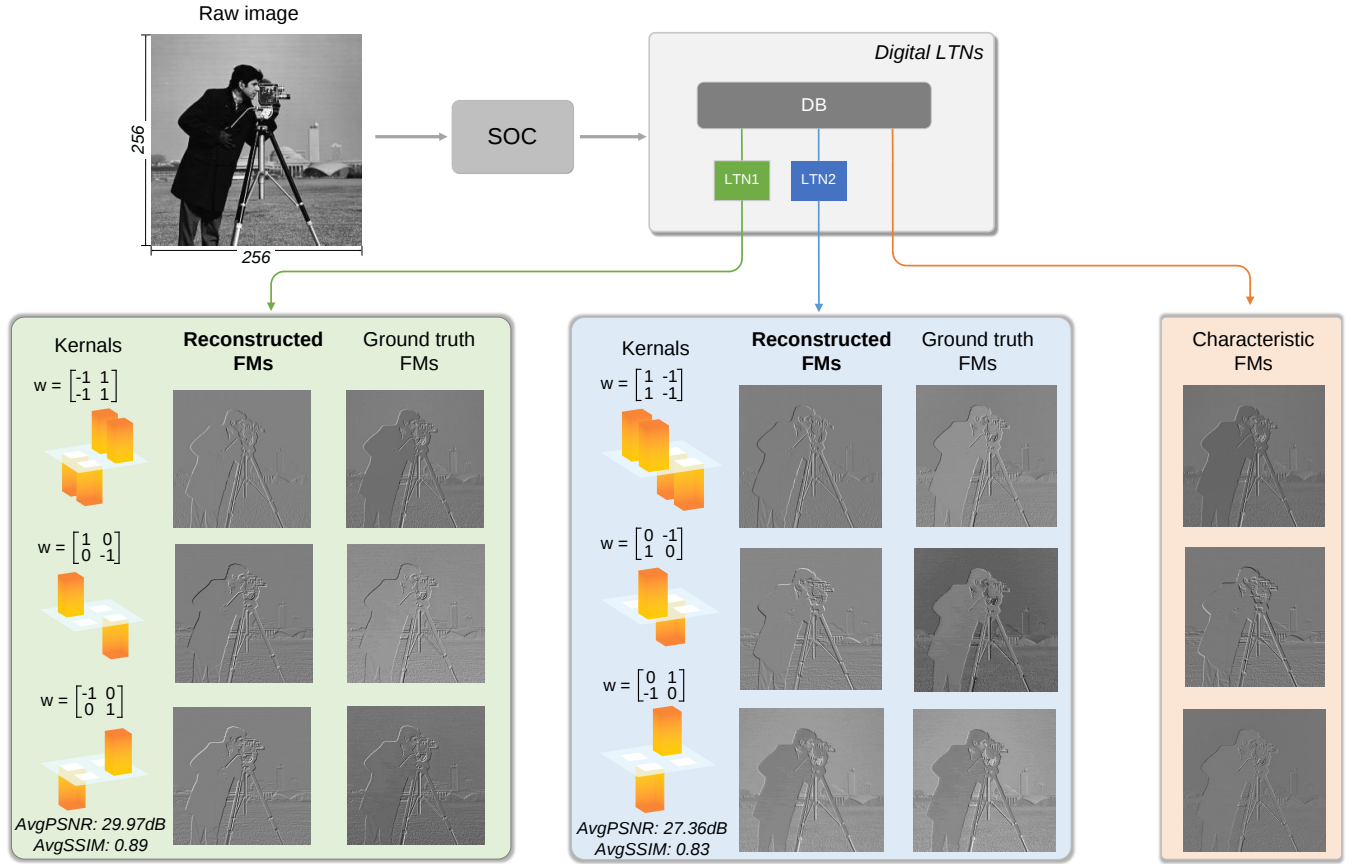


Figure S5: Another experimental image and the outputs of specific reconstructed convolution kernels and the characteristic kernels of the SOC, compared with the ground truth from commercial 64-bit computer.

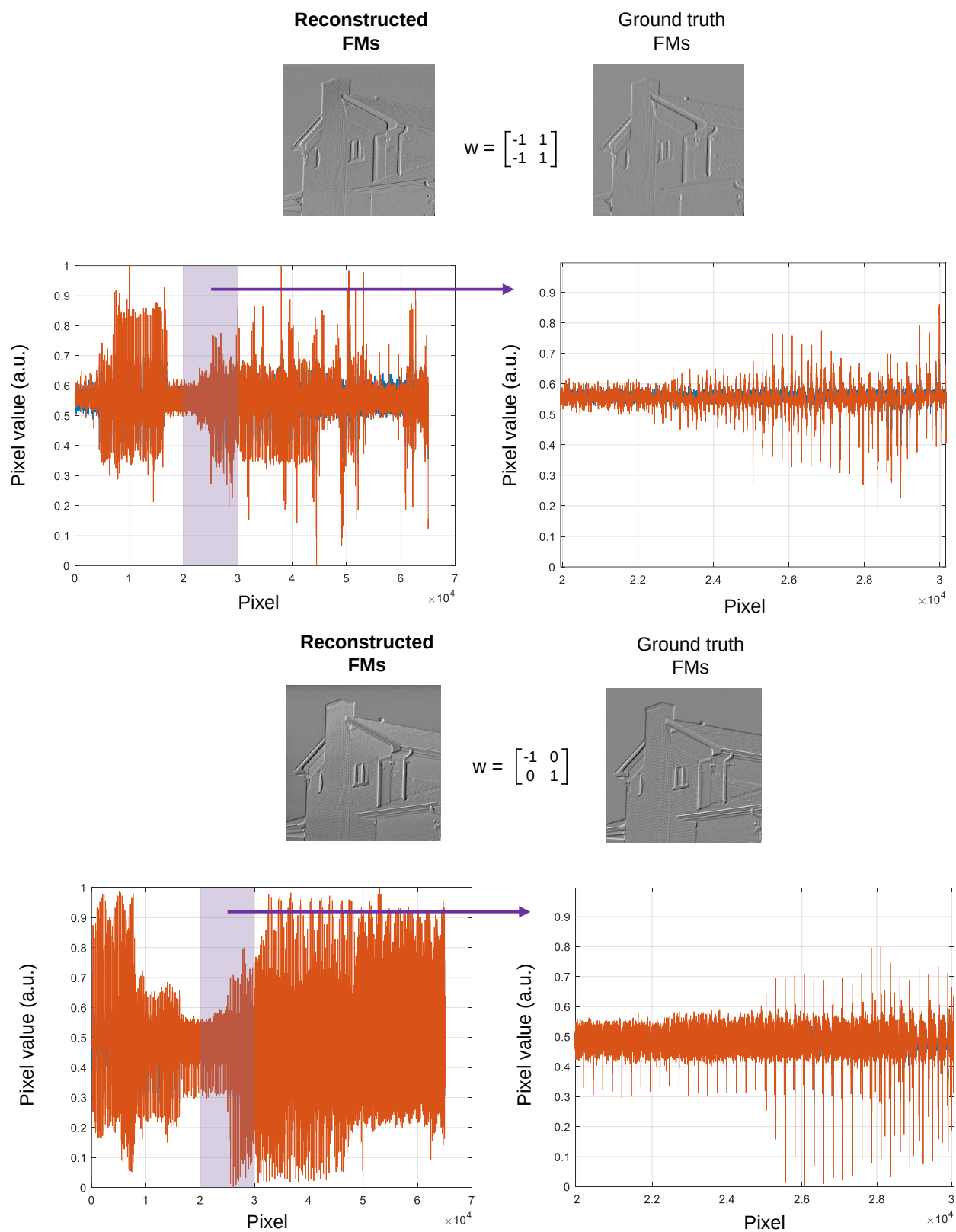


Figure S6: Comparison of serialized pixels of the two produced feature map. Blue line, experimental results; Orange line, simulation results.

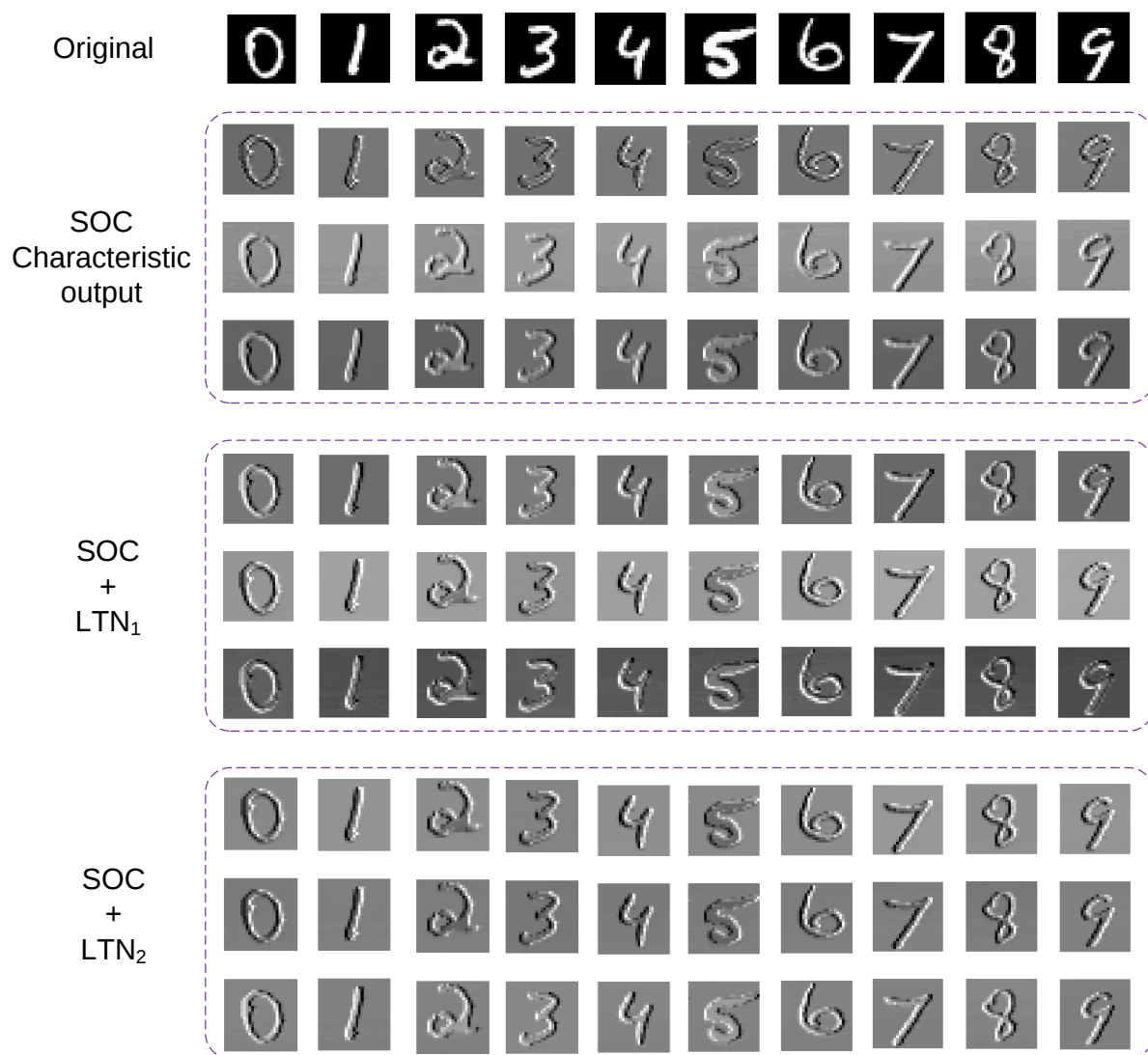


Figure S7: Experimental results of MNIST handwritten digits feature maps produced by SOC, SOC + LTN₁ and SOC + LTN₂, respectively.

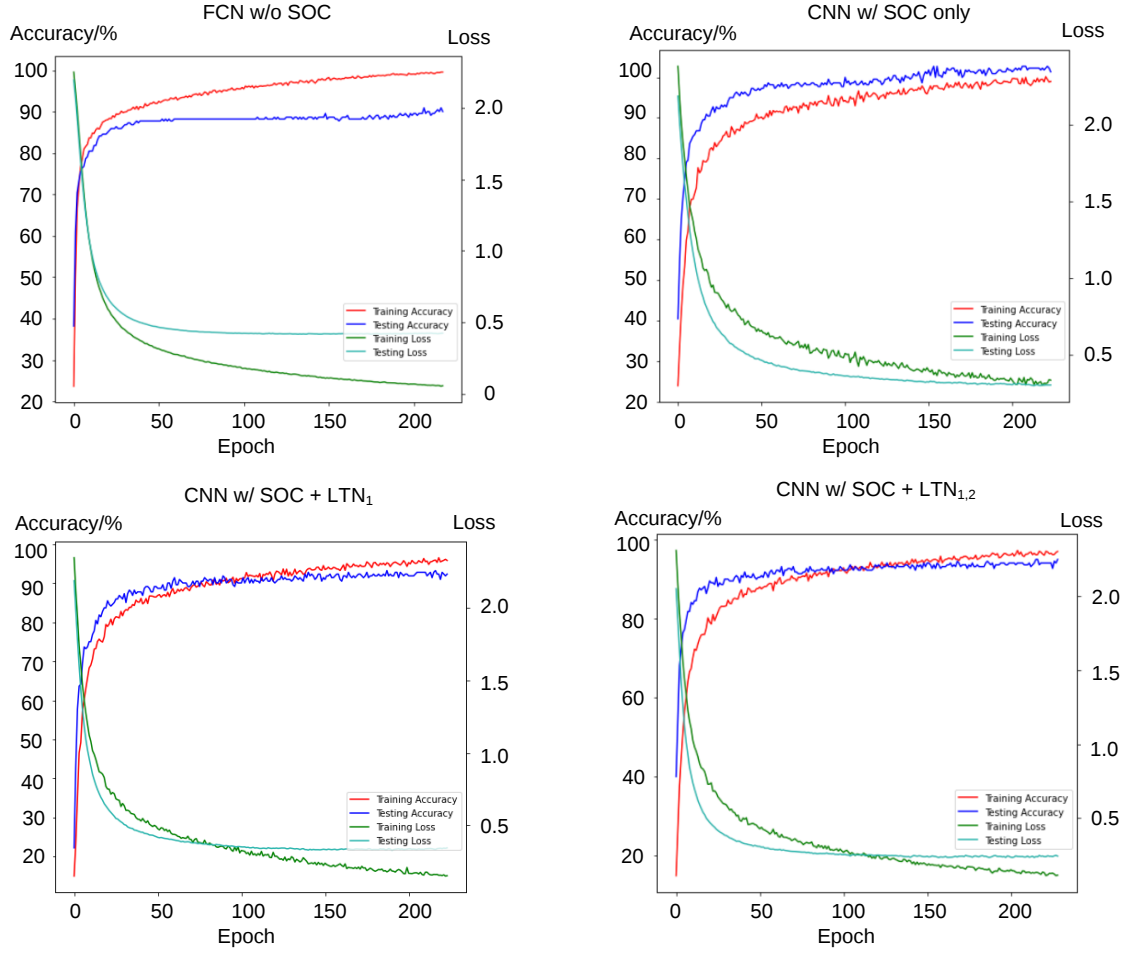


Figure S8: The accuracy and loss curves of MNIST handwritten digits dataset for four network structures. The training dataset includes 2,000 images and the testing dataset includes 200 images.

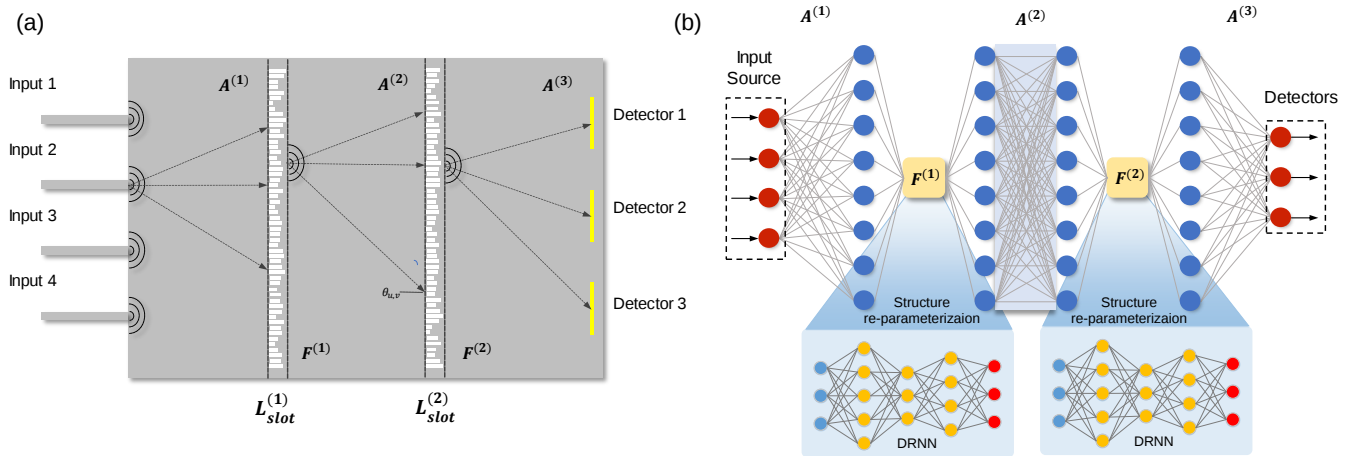


Figure S9: (a) Schematic of proposed SOC, each point on a given layer acting as a secondary source of a wave, the on-chip propagation in different area can be calculated by matrix A and function F respectively. (b) Network abstraction of forward propagation process. The matrix A can be represented by Huygens Fresnel principle and function F can be mapped to DRNN in Supplementary note 1.

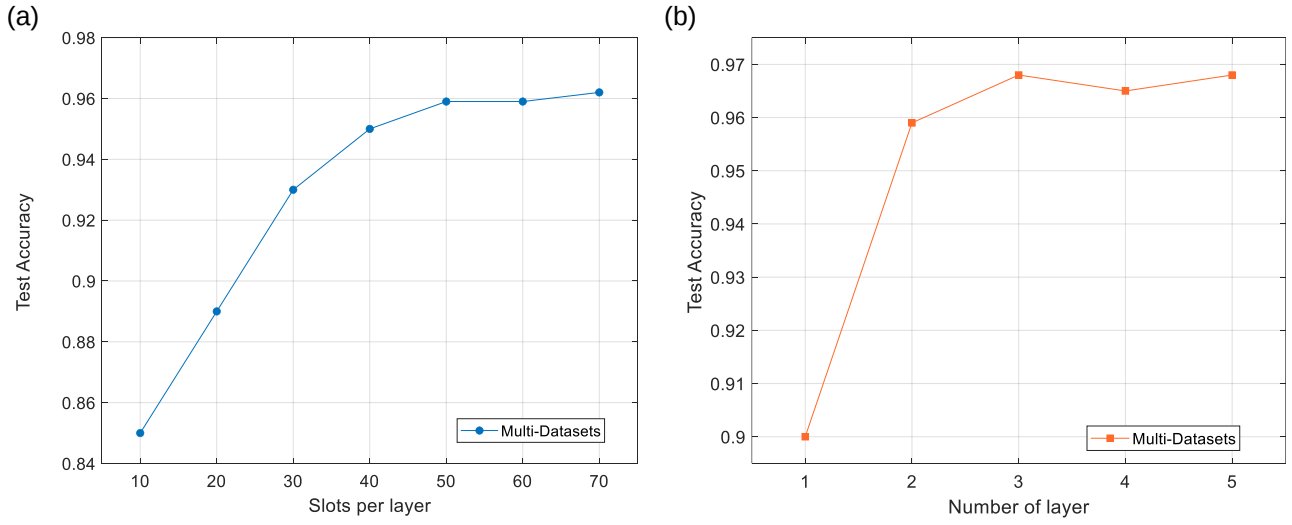


Figure S10: (a) Impact of the number of slots per layer on the classification accuracy. (b) Impact of the number of layers on the classification accuracy.

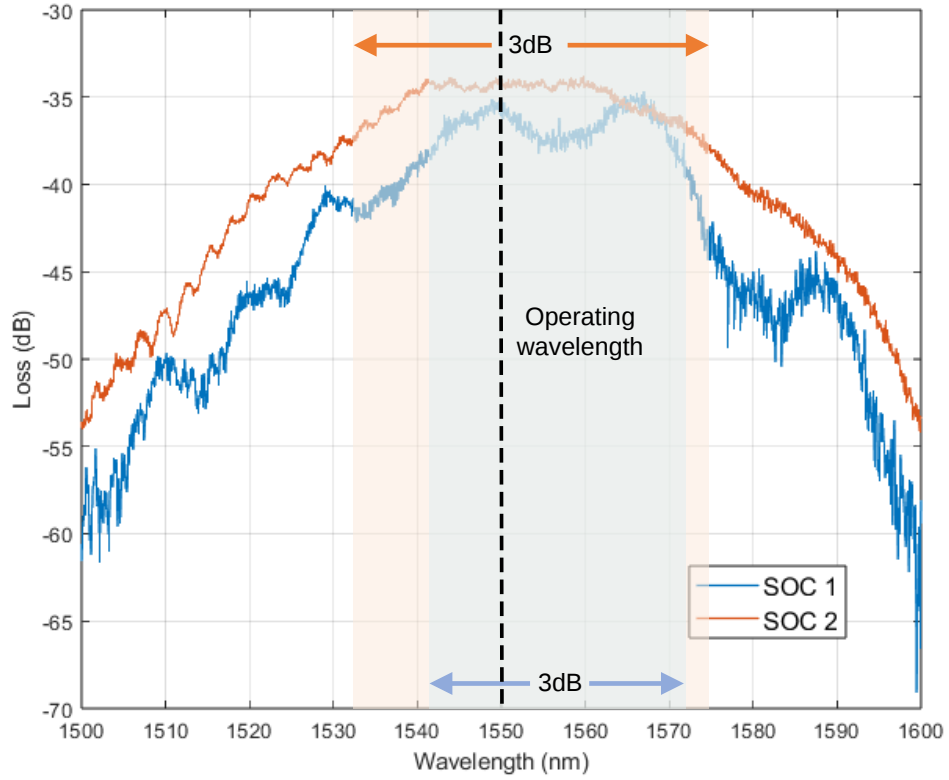


Figure S11: Experimental results of wavelength-scan of the two SOC chips.

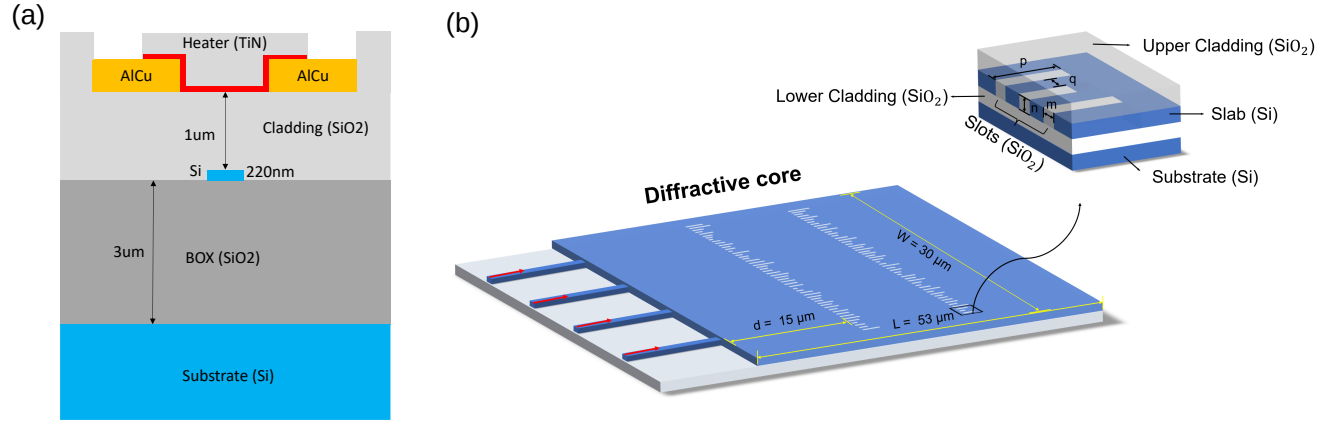


Figure S12: (a) The fabrication process details of our designed SOC chip. (b) Structure of the proposed SOC chip.

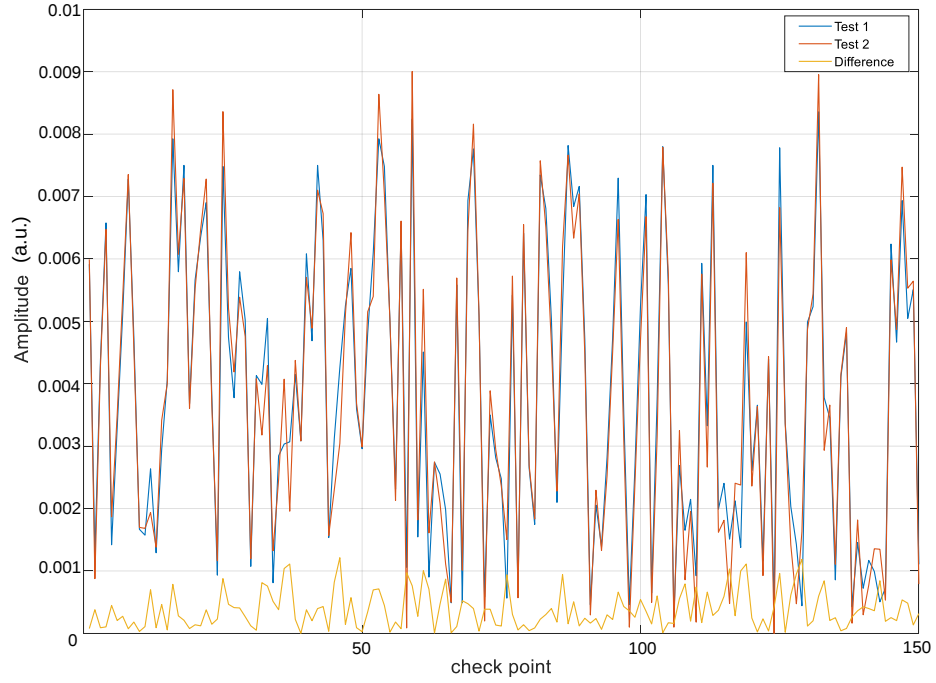


Figure S13: Comparison of test results for the same chip at two different times and under reconnection conditions.

References

- [1] Tingzhao Fu, Yubin Zang, Yuyao Huang, Zhenmin Du, Honghao Huang, Chengyang Hu, Minghua Chen, Sigang Yang, and Hongwei Chen. Photonic machine learning with on-chip diffractive optics. *Nature Communications*, 14(1):70, 2023.

- 165 [2] Yuyao Huang, Tingzhao Fu, Honghao Huang, Sigang Yang, and Hongwei Chen. Sophisticated deep
166 learning with on-chip optical diffractive tensor processing. *Photonics Research*, 11(6):1125–1138, 2023.
- 167 [3] Tingzhao Fu, Yubin Zang, Honghao Huang, Zhenmin Du, Chengyang Hu, Minghua Chen, Sigang
168 Yang, and Hongwei Chen. On-chip photonic diffractive optical neural network based on a spatial
169 domain electromagnetic propagation model. *Optics Express*, 29(20):31924–31940, 2021.
- 170 [4] Steven H Strogatz. Exploring complex networks. *nature*, 410(6825):268–276, 2001.
- 171 [5] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe
172 Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex
173 networks. *arXiv preprint arXiv:1705.09792*, 2017.
- 174 [6] Maxime W Matthès, Yaron Bromberg, Julien de Rosny, and Sébastien M Popoff. Learning and
175 avoiding disorder in multimode fibers. *Physical Review X*, 11(2):021060, 2021.
- 176 [7] Wencan Liu, Tingzhao Fu, Yuyao Huang, Run Sun, Sigang Yang, and Hongwei Chen. C-donn: com-
177 pact diffractive optical neural network with deep learning regression. *Optics Express*, 31(13):22127–
178 22143, 2023.
- 179 [8] Yuyao Huang, Wencan Liu, Run Sun, Tingzhao Fu, Yaode Wang, Zheng Huang, Sigang Yang, and
180 Hongwei Chen. Diffraction-driven parallel convolution processing with integrated photonics. *Laser &
181 Photonics Reviews*, page 2400972, 2024.
- 182 [9] Zhihao Xu, Tiankuang Zhou, Muzhou Ma, ChenChen Deng, Qionghai Dai, and Lu Fang. Large-scale
183 photonic chiplet taichi empowers 160-tops/w artificial general intelligence. *Science*, 384(6692):202–209,
184 2024.