Programmable Surface Plasmonic Neural Networks

Xinxin Gao

Southeast University

Qian Ma

Southeast University   https://orcid.org/0000-0002-4662-8667

Ze Gu

Southeast University

Wen Yi Cui

Southeast University

Che Liu

Southeast University   https://orcid.org/0000-0002-9917-8487

Jingjing Zhang

Southeast University

Tie Jun Cui (✉ tjcui@seu.edu.cn)

Southeast University   https://orcid.org/0000-0002-5862-1497

Article

Keywords:

Abstract

Recently, some new forms of artificial intelligence computing hardware and chips have been presented. However, most of them have difficulties to simultaneously achieve advantages of light-speed computing, programmable weight matrix, and programmable nonlinear activation functions. Here, we propose a programmable surface plasmonic neural network (SPNN) with programmable weights and activation functions based on a spoof surface plasmon polariton (SSPP) platform, which can perform intelligent functions and sense electromagnetic (EM) waves at nearly light speed. We demonstrate a parallel coupling SSPP structure loaded with varactors to introduce four paths with tunable transmitting parameters. On this multi-port architecture, we further establish a real-time control and feedback method to enable arbitrarily designable activation functions under a detecting feedback loop. Experimental results show that a four-in and four-out fully-connected super-neuron can fulfill independently adjustable weights and programmable activation functions, where each input can be sensed for arbitrarily programming. To comprehensively show the above capabilities, we design and demonstrate experimentally a wireless communication system based on the SPNN for image decoding and recovery. We further illustrate a partially connected SPNN using the super-neurons with a high prediction accuracy. The proposed concept paves a new way for artificial intelligence devices, stimulating the fascinating fields like large-scale EM computing and communication systems in the future.

Introduction

Artificial neural networks (ANNs) are computational network models inspired by the signal processing in human brain,[1, 2] which can forecast the correlation between a set of input and output parameters by simulating the principles of biological neural networks (NNs).[3] The development of ANNs that could rival their biological counterparts is arguably regarded as the last frontier in computing. [4] ANNs have achieved outstanding performance ranging from the computer vision to speech recognition, natural language processing, image denoising, chemistry, and medicine.[5−11] As the demands for brain science and artificial intelligence (AI) grow, more complicated problems and high calculation accuracy have requirements for deep neural networks (DNNs).

Initially, the architecture of DNNs is devoted to the research of network algorithms, including convolutional neural networks and recurrent neural networks. [5−8, 12, 13] Recently, to improve the calculation speed and reduce power dissipation, different AI circuits and chips have been designed and even commercialized. The nanophotonic circuit as a promising alternative has also been used extensively in ANNs. [1, 14−20] These optical neuron networks utilize integrated optical waveguides and active optical devices (e.g. phase modulators) to perform the matrix multiplication. Thereafter, a fully hardware-implemented memristor AI chip with high energy efficiency was proposed, which integrates eight 2048-cell memristor arrays to improve the parallel-computing efficiency. [21] However, the AI chip is only suitable for handling analog signals and unable to sense the electromagnetic (EM) waves. Besides, the calculation speed of the AI chip is limited. Recently, an innovative all-optical diffractive deep neural

network (D^2NN) architecture, whose foundation lies on the Huygens's principle and tunable phase transition at each neuron, has the capabilities to calculate at light speed and manipulate the EM waves. [11, 22] Inspired by this concept, a reprogrammable D^2NN based on multi-layer information metasurfaces has been reported. [23] The reprogrammable D^2NN could not only modulate the EM waves for intelligent computing at the light speed, but also has great potentials in wireless communications. However, these proposed D^2NNs are the linear model without nonlinear activation functions, limiting their application potentials. It is well known that the nonlinear activation functions play an important role in ANNs by enabling them to learn complex mappings between inputs and outputs and compensate for the shortage of the linear models. [24]

Spoof surface plasmon polariton (SSPP) waveguide could be considered as an alternative technology for achieving ANNs, due to the advantages of low crosstalk, low radiation loss, and easy integration. [25] SSPPs are surface waves in the microwave and terahertz frequencies, mimicking the unique properties of natural surface plasmon polaritons (SPPs) via constructing structured metal surfaces. [26] Hence, SSPPs have strong field enhancements and can freely control the EM waves by manipulating the dispersion behaviors. [27] Subsequently, an ultrathin and flexible SSPP waveguide was proposed to achieve the miniaturization and integration of plasmonic devices. [28] Follow-up research found that the SSPP signals can be transmitted in a highly dense network, in which the crosstalk between channels is effectively suppressed. [29–32] Some SSPP devices have been reported, such as couplers, power dividers, and filters. [33–35] Especially in the SSPP coupler, the SSPP signals can be freely controlled in real-time by altering the bias voltages of varactors loaded on the SSPP waveguides. [33] In terms of these merits, the SSPP waveguides have great potentials to realize the interconnections of a large-scale and reprogrammable ANN.

Here, we propose and experimentally demonstrate a programmable surface plasmonic neural network (SPNN) composed of SSPP super-neurons, which can not only achieve the flexible weight modulations and designable activation functions but also perform image encoding and decoding, working in microwave frequencies. Each SSPP neuron includes two programmable SSPP couplers and one SSPP power divider, which exactly establishes the relationship among inputs as a typical neuron node. The weight coefficients can be regulated flexibly by tuning the voltages of varactors integrated on the SSPP couplers. Moreover, the activation function of the created programmable SPNN can be fully customized by detecting the input intensity from detectors and feeding back the threshold to an amplifier. Handwritten digits are successfully recognized by training the partially connected physical SPNN, which has better training performance than the fully connected network with the same scale. To exhibit the direct sensing and computing abilities of SPNN in the EM space, we demonstrate experimentally image encoding/decoding and recovery for the wireless communications based on the SPNN without using any conventional decoding devices.

Results And Discussions

Design, fabrication, and measurement of programmable SSPP neuron. Fully connected or partially connected ANNs usually require each neuron to have multi-input and multi-output, as well as flexible and independent weight modulations. Here, we present a physical SPNN with a partially connected network architecture, as shown in Fig. 1. The proposed SPNN is accomplished in layer-by-layer fashion, where each layer consists of massive programmable SSPP super-neurons. Each super-neuron with a four-in and four-out fully connected network is composed of eight programmable SSPP neurons. As an essential unit, the programmable SSPP neuron with four inputs and one output has the following three characteristics: 1) high-efficiency SSPP interconnects, 2) programmable weight modulations, and 3) programmable activation functions. Benefiting from the powerful EM manipulations and easy integration of the SSPP structures, we could realize programmable transmission weight modulations by embedding varactors on a coupling waveguide structure. Simultaneously, we introduce two RF detector chips in each programmable neuron for real-time monitoring of the input. Such a mechanism will enable an arbitrarily programmable threshold design for various input distributions. Then, the collected information generated by the detector is fed back to the power amplifier chip used as the active component for a designable transmission response. In this process, a customized activation function will be actualized. Finally, the weighted inputs are combined to enter a single waveguide using a power divider and are guided to the output via the power amplifier.

To realize the above idea, we design the programmable SSPP neuron that can realize the weighted modulations and implement the SPNN to control arbitrary activation functions. To build the relationship between the input and output information and further realize the programmable weight modulations, two different SSPP unit structures are presented, as seen in Fig. 2a, in which the structural parameters are a = 2 mm, d = 0.75 mm, p = 2 mm, and s = 0.25 mm. The structure is printed on the FR4 substrate (ε_r = 4.3 and $\tan\delta$ = 0.025) with a thickness of 0.8 mm. The dispersion behaviors are simulated by the Eigen-mode solver, CST Microwave Studio, as sketched in Figs. 2b and 2c. We note that the SSPP mode moves gradually away from the light line when the frequency increases, mimicking the dispersion behaviors of optical SPPs. For the unit shown in the left-upper corner in Fig. 2a, we observe that the cutoff frequency gradually decreases as the groove depth (h) increases from 1.25 to 3.25 mm, as shown in Fig. 2b, indicating that SSPPs begin to exhibit stronger field confinements. In particular, when h = 2.25 mm, the dispersion relations of the mirror-symmetric SSPP unit integrated with the capacitors (see the right-lower corner in Fig. 2a) are illustrated in Fig. 2c. We see that the phase constant at the same frequency gradually increases as the capacitance varies from 0.03 to 0.25 pF, denoting that the EM fields are progressively concentrated from both sides of the waveguide to between the waveguides. Hence, the SSPP signals between the waveguides can be freely distributed by manipulating their dispersion behaviors.

The programmable SSPP output neuron consisting of the SSPP units mentioned above is shown in Fig. S1, where two parallel SSPP couplers loaded with tunable varactors (MAVR − 011020 − 1411), an SSPP power divider, a power amplifier chip (Qorvo TQP369180) [36], and two RF detector chips (LTC5530) [37] are integrated on this neuron. The detailed analysis is provided in Supplementary Information 1. The two-

parallel programmable SSPP couplers and SSPP power divider can build the relationship between the input and output signal waves by changing the capacitance of varactor. The continuous state transition is illustrated in Fig. 2d, where the relationship between two S-parameters can be represented by a linear function. The operating frequency of 4.6 GHz is selected, where the linearity and the dynamic ranges reach the best performance.

To verify the effectiveness of the above simulation, we fabricated and measured a sample of the programmable SSPP output neuron (see Fig. 3a and Methods). As shown in Fig. 3b, the coupling transmission decreases gradually and direct-through transmission increases progressively as the bias voltages applied to varactors vary from 0 to 15 V, in which the experimental results are consistent with the simulated results presented in Fig. 2d. Thus, we can adjust the bias voltages applied to varactors by using a digital-to-analog converter (DAC) and further achieve the programmable weight modulations.

Additionally, at a series of incident intensities varying from 0 to 23 dBm, the scattered output spectra are measured and the monitoring performance of the SSPP neuron is verified (see Supplementary Information 2). We clearly observe that the waveform of sensing ports can be controlled by switching the bias voltage of varactors, as shown in Fig. S5, indicating that the detector can monitor the inputs in real-time. Figure 3c also shows the relationship between direct-through and coupling parameters as the bias voltage of the amplifier changes. Hence, customized activation functions in the SSPP neuron can be achieved by the feedback loop between the detector and amplifier. The implementation process can be summarized as the following steps: 1) setting the bias voltages applied to the varactors; 2) detecting the input power intensity via the detector; 3) switching the bias voltages applied to the varactors; 4) detecting the input information again; 5) judging the collected information; 6) enabling the amplifier according to the pre-loaded activation functions. Finally, when the detected output voltage generated from the inputs feeds back to the amplifier, a customized activation function will be realized (see Supplementary Information 5). Figures 3d to 3g depict the realization of four classic types of activation functions (see Methods). We remark that certain power degradation is observed as the input power or the voltage for the amplifier increases, which could be compensated in the feedback loop.

SSPP super-neuron and encoding-decoding wireless communication. In terms of the good characteristics of the programmable SSPP neuron, we fabricate a super-neuron, as shown in Fig. 4a, mimicking the simplest two-layer fully-connected SPNN with four-in and four-out. Note that the SSPP super-neuron consists of four input neurons and four out neurons, in which the input neurons are not loaded with the amplifiers but the output neurons are. This interlocking structure enables the realization of a customized weight matrix (Fig. 4b), which is derived numerically (see Methods). We firstly validate the proposed SPNN on a customized vector classification task. Here, each category of the dataset is defined by the addition of a basis vector with Gaussian white noise to resemble the received signal intensity of multi-channel amplitude shift keying (ASK) modulation in the communication systems. By using a classical back-propagation algorithm and the ReLU activation function to train the proposed SPNN, the ideal weight matrix can be computed over the testing dataset, as shown in Fig. 4c. Additionally, the confusion matrix, prediction accuracy, and cross-entropy loss are also computed, as presented in Fig. S9.

To verify the effectiveness of the ideal result, we measure the transmission of the SSPP super-neuron, and the measured data are fitted to obtain the theoretical matrix (Fig. 4d). We observe that the ideal and theoretical matrices are in excellent agreement. Finally, the experimental measurement platform is set up (see Supplementary Information 6) to measure the four-in and four-out super-neuron. By training the super-neuron model, the experimental weight matrix is achieved, as illustrated in Fig. 4e. The corresponding experimental and theoretical results keep a good consistency. We remark that, in order to achieve high accuracy of the proposed SPNN, we verify the consistency of four input neurons and four output neurons (see Supplementary Information 4). The predicted accuracy based on the theoretical and experimental weight matrices is given in Fig. 4f. Compared to the theoretical result, the experimental result has small performance degradation due to the simulation errors and other factors. However, it correctly reflects the original weight matrix. On balance, the high training performance of the SSPP super-neuron is verified by experiments.

To demonstrate the programmable functionality of the proposed SPNN, we further build an encoding/decoding wireless communication system by replacing the conventional baseband processing unit with SPNN, as shown in Fig. 5a. To realize the data-stream transmissions, the transmitting end of the designed communication system firstly encodes the binary data into symbols (see Code 0 to Code 3 in Fig. 5d), and then maps the symbol streams into the ON/OFF states of signals from four transmitted antennas through RF switches. Then the propagated EM fields are collected by four receiving antennas. The estimated channel matrix between the Tx/Rx antennas, inferred from the simulation results in Fig. 5c, could be assumed as an identity matrix. Therefore, the signal amplitudes of four channels in the receiving end share a similar code mapping as that in the transmitting end. At the receiving end, decoding is carried out to restore the raw bit-streams from the received RF signals.

In the traditional wireless communication system, the conversion from microwaves to baseband signals is always performed directly on the received RF signals, without complex operations in the EM domain. In the proposed communication configurations, however, the signals from the transmitted antenna array are directly fed into the SPNN for signal processing in the EM domain, as demonstrated in Fig. 5a. Figure 5b depicts an abridged diagram illustrating the decoded signals at the four outputs of the SSPP super-neuron. The amplitudes of the signals are converted to an injective mapping related to the original code pattern: the maximum at Port 1 corresponds to Code 0, the maximum at Port 2 corresponds to Code 1, the maximum at Port 3 corresponds to Code 2, and the maximum at Port 4 corresponds to Code 3. As a result, the data stream could be decoded with little effort. In the experiments, each output port of the SSPP super-neuron is connected with an RF detector, followed by an ADC module for coding detection.

To verify the capability of the proposed communication system, a 230×230-pixel logo picture of Southeast University is transmitted, which is represented by a binary sequence of 52900 bits or a 2-bit sequence of 26450 symbols. Two insets in Fig. 5a illustrate the original picture (the right-upper corner) and the decoded picture (the left-upper corner). With a total error rate of 0.04%, the reconstructed picture exhibits negligible deterioration of Fig. quality, confirming the functionality of the system. The transmission rate of the system is 4120 bits per second or 2060 symbols per second, which is currently

limited by the ADC sampling rate in our experiments. We remark that the chosen modulation codes shown in Fig. 5d are only determined in this case for the consistency of the SPNN states in Figs. 4c-4e. In reality, arbitrary symbol mapping could be applied, as long as the corresponding decoding procedure is programmed into the SPNN. It could also be inferred that the SSPP super-neuron could realize the compensation for the channels and other components in the analog front ends in modern wireless communication systems. More importantly, in the proposed communication system, the decoding rate would no longer be constrained by the baseband processing clock rate, as the procedure is now carried out in our SSPP super-neuron module in the EM domain, which is theoretically performed at approximately the speed of light.

Partially connected SPNN for automatic picture identifications. To establish a deeper and more complex network with more SSPP super-neurons, a partially connected SPNN structure is designed and constructed, as depicted in Fig. 6a. Here, an SSPP super-neuron can be regarded as a typical neuron node. Although the SSPP super-neuron is a fully connected network, the connecting relationship among the SSPP super-neurons is a partially connected network. The SPNN depicted in Fig. 6a is composed of an input layer including four super-neurons (16 inputs), four hidden layers, and an output layer (10 outputs). To fit the size of the input vectors, discrete-cosine transformation (DCT) is carried out on each picture sample in the MNIST dataset, and the first 16 main components from the DCT results are extracted as a new set of the input training data. [24] Different from the weight matrix of the fully connected ANN, in which each element of the 16x16 weight matrix is adjustable, only a limited number of the weight elements will be trained in the partially connected SPNN (see Supplementary Information 8). By training this network with the ReLU activation function and the modified back-propagation algorithm, the confusion matrix is calculated, as shown in Fig. 6b. We clearly observe that the partially connected SPNN can successfully classify handwritten digits. When the number of input neurons increases from 16 to 32, the classification ability is improved by comparing the obtained confusion matrices, as demonstrated in Fig. 6c. In particular, Figs. 6d and 6e depict the classification rates of the 16 and 32 input neurons, respectively. For each network, 64 parallel simulations are carried out on the same database.

To set up a fair comparison, we confine a fully connected network with the same number of variable weight factors as the counterpart (see Supplementary Information 9). Comparing the orange and purple bars in Figs. 6d and 6e, we deduce that the partially connected network can achieve higher accuracy than the fully connected network in the presented scale, for the networks with the 16 and 32 input neurons, respectively. Therefore, a practical SPNN structure can provide excellent performance and could be physically implemented using the proposed SSPP super-neuron structures.

Conclusion

We demonstrate that high-performance learning has been realized on the SSPP platform by training the proposed reprogrammable SPNN, holding promising for on-chip integration with the microwave circuits. Owing to the flexibility of the SSPP dispersive behaviors, arbitrary weight matrix modulations are achieved by altering the bias voltages applied to the varactors on the SSPP structures. Our SPNN can

implant the activation function into the SSPP circuits with customized functions through a feedback loop between the detector and power amplifier in each neuron. The proposed SPNN can directly sense and process the EM waves and further implement the image encoding and decoding in the wireless communication systems. The concept of partially connected SPNN is firstly proposed in this work, which has comparable training performance with its fully connected counterpart. We believe that this work will extend the scope of plasmonic metamaterials and establish a foundation for novel artificial intelligence engineering to facilitate the fast and reliable machine learning processes.

Methods

Numeric model of SSPP super-neurons. Each SSPP super-neuron consists of eight neurons, including four inputs and four outputs, as shown in Figs. 1 and 4a. For the output neurons, two RF detectors and an amplifier are integrated to monitor the input and realize the programmable activation functions. For the input neurons, the detector and amplifier chips are omitted and replaced by matching loads and microstrip lines, respectively, because the input neurons do not involve the realization of the nonlinear part. As to be shown in later numeric derivations, this interlocking structure enables the realization of a customized weight matrix.

If we mark the input signal of four input ports as x and that of the four output ports as y, the proposed programmable SSPP super-neuron could theoretically realize arbitrary matrix multiplication between x and y. That is to say, given any weight matrix $\mathbf{W}$ of size 4×4, we could set the bias voltages applied to the varactors accordingly in our SSPP super-neuron to satisfy

$$y = \mathbf{W}x \qquad (1)$$

We will now derive the relationship between the weight (transmission) matrix of the SSPP super-neuron and the coupling factor. Firstly, we explore the numeric model of a single coupler, which is the basic component of the whole network. For an ideal coupler with a changeable division ratio, the square magnitude of the signal at direct through and coupling ports could be written as

$$|S_{couple}|^2 = \alpha \tag{2}$$

$$|S_{direct}|^2 = 1 - \alpha \tag{3}$$

or

$$|S_{direct}|^2 = 1 - |S_{couple}|^2 \tag{4}$$

In the real model, the equation is modified to

$$|S_{direct}|^2 = g_0 - k_0 |S_{couple}|^2 \tag{5}$$

where g_0 (intercept) represents the total energy loss from the path loss on dielectric substrate and resistance of varactor, and k_0 (slope) represents a constant describing the loss ratio of the coupling signal due to the parasitic resistance of varactors. Based on physical interpretation, g_0 should be a value smaller than 1, while k_0 is a value greater than 1 (assume no amplifier). The simulated result in Fig. 2d confirms the inference, with $g_0 = 0.294$ and $k_0 = 1.773$. In fact, the ideal reference of g_0 should be 0.5 in this case due to the existence of a power divider at the front of the circuit. This together means a power loss of around 60% (2.2 dB) for one SSPP neuron board, which could undoubtfully be alleviated by selecting a low loss-tangent substrate. Meanwhile, the R-square of the linear fitting for the simulated result in Fig. 2d reaches up to 0.987, proving that a simple quasi-linear model is adequate to depict the performance of a single power divider. We remark that g_0 and k_0 are determined for each SSPP board based on the measurement and linear fitting to guarantee a minimum deviation from the model.

Now, we consider the model for four input neuron boards, each etching two couplers. Without the loss of generality, we neglect the imperfect reflection loss of each port and crosstalk between different structures. Then, we can derive the transmission parameters of the input neuron board as

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{21} & u_{31} & u_{41} \\ g_0 - k_0 u_{11} & g_0 - k_0 u_{21} & g_0 - k_0 u_{31} & g_0 - k_0 u_{41} \\ g_1 - k_0 u_{12} & g_0 - k_0 u_{22} & g_0 - k_0 u_{32} & g_0 - k_0 u_{42} \\ u_{12} & u_{22} & u_{32} & u_{42} \end{bmatrix}$$

6

The sixteen-element matrix could be considered as the intensity of 16 intermediate ports when the four input ports are injected with normalized power, i.e., transmission parameters between the input and 16 intermediate ports. Each column of **U** represents four outputs of a single board, and u_{i1} and u_{i2} ($i = 1, 2, 3,$

4) stand for the signal received at the coupling ports, where the signals at the direct ports could be derived by Eq. (5). Moreover, in Supplementary Information 5, we exhibit the measurement results of the consistency of the fitting numbers of g and k, which are mainly determined by the fabrication accuracy.

Back to the SSPP super-neuron model, we similarly depict the transmission parameters of the output neuron boards as (neglect the amplifier at the output terminal temporarily)

$$\mathbf{K} = \begin{bmatrix} k_{11} & g_1 - k_1 k_{11} & g_1 - k_1 k_{12} & k_{12} \\ k_{21} & g_1 - k_1 k_{21} & g_1 - k_1 k_{22} & k_{22} \\ k_{31} & g_1 - k_1 k_{31} & g_1 - k_1 k_{32} & k_{32} \\ k_{41} & g_1 - k_1 k_{41} & g_1 - k_1 k_{42} & k_{42} \end{bmatrix}$$

7

Each row of represents the transmission parameters from four input ports in the intermediate layer to the output port in a single output-neuron board. Again, g and k are considered to be the same for the four boards.

Finally, the amplifier is represented as a diagonal matrix $\mathbf{B}$ = diag (B_i) (i = 1, 2, 3, 4), and B_i represents the transmission factor between the input and output of the ith amplifier. The nonlinear characteristic of the network is based on the controlled loop B_i. Here we could temporarily assume B_i as 1. Then, the weight matrix in Eq. (1) could be determined by

$$\mathbf{W} = \mathbf{B}\left(\mathbf{K} \odot \mathbf{U}\right) = \left(\mathbf{B}\mathbf{K}\right) \odot \mathbf{U}$$

8

where represents the element-wise product, whose physical meaning is the concatenation of the input-neurons $\mathbf{U}$ and output-neurons $\mathbf{K}$ in the 16 intermediate ports.

Given an arbitrary 4×4 matrix $\mathbf{W}$, we could determine $\mathbf{U}$ and $\mathbf{K}$ respectively from Eq. (8) while assuming that $\mathbf{B}$ is a unitary matrix. The core is to solve four independent quartic equation sets. We remark here that different g and k in $\mathbf{U}$ and $\mathbf{K}$ (as in equations (6) and (7)) merely influence the coefficients in the equations and have little impact on the fore-mentioned process. Apparently, $\mathbf{W}$ has a freedom degree of 16, and $\mathbf{U}$ and $\mathbf{K}$ share a freedom degree of 8 (u_{ij}, k_{ij}, i = 1, 2, 3, 4, j = 1, 2). Consequently, arbitrary $\mathbf{W}$ could be realized based on the proposed SSPP super-neuron structure. However, the equation will sometimes have complex solutions or the solution goes beyond the dynamic range of the power coupler. In such circumstances, we could minorly alter the amplifier factor B_i to find the optimum realizable solutions or simply limit the out-of-range solutions within the admissible range, while introducing some errors at the same time. In our experiments, the elements in the weight matrix are limited to a specific boundary during the training process.

Voltage control of varactors and amplifiers. The bias voltage controlling the capacitance of varactors reaches up to 12 V, while the bias voltage resolution of amplifiers requires to be less than 10 mV. Therefore, the DAC chip should provide a high voltage range for the adequate resolution bit together with multiple channel selection. Based on these requirements, AD5371 (Analog Cooperation) chip is chosen as the multi-channel voltage generator. Sixteen channels control the bias voltages of varactors in the SSPP super-neurons. Four channels connect to the output pins of the four amplifier chips through inductors. The controlling command for voltage setting is sent from the microcontroller unit (MCU) through the serial peripheral interface (SPI). We note that the command bytes for the varactor bias voltages are sent once at the beginning, while the command bytes for the amplifier voltages may be sent multiple times at different levels based on the detected input intensity to realize the nonlinear functionality.

Measurement of nonlinear characteristics. The vector network analyzer (Agilent N5230C) functions as an S-parameter measurement instrument and a programmable power generator. The input power is programmed as 47 scales ranging from 0.25 mW to 6 mW, where the corresponding output power is calculated by the multiplication of the measured transmission parameters. We simplify the measurement procedures by injecting a single port in the neuron board, while the other three ports are connected to match the loads.

Measurement of neurons and programmable super-neurons. We use the vector network analyzer to measure the coupling and direct-through transmission parameters by switching the bias voltage applied to the varactors and changing the voltage of the power amplifier. We note that all ports are connected to the 50 Ω matched loads to avoid reflection, except for the test input and output signal ports. The operation process is controlled by the digital-to-analog converter (DAC). Additionally, to obtain the activation function, it is necessary to introduce an analog-to-digital converter (ADC). The detector collects the input information to ADC, which transports the information to MCU. Then the collected signal by MCU is sent to DAC. Finally, the amplifier obtains the control voltage by DAC.

Declarations

Acknowledgments

References

1. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. Nat. Photon. **11**, 441–446(2017).
2. Krogh, A. What are artificial neural networks? Nat. Biotechnol. **26**, 195–197(2008).
3. Huang, H. M. et al. Artificial Neural works based on memristive devices: from device to system. Adv. Intellig. Systems **2**, 2000149(2020).

4. Strukov, D. B. Smart connections. Nature 476, 403–405(2011).

5. Tompson, J. et al. Joint training of a convolutional network and a graphical model for human pose estimation. Adv. Neural Inf. Process. Syst. 1–9(2014).

6. Graves, A. Mohamed, A. Hinton, G. Speech recognition with deep recurrent neural networks. *in Proceedings of the* 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*. 6646–6649(2013).

7. Young, T. et al. Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**, 55–75(2018).

8. Xie, J. Xu, L. Chen, E. Image denoising and inpainting with deep neural networks. Adv. Neural Inf. Process. Syst. (1–9)2012.

9. Gilmer, J. et al. Neural message passing for quantum chemistry. *in Proceedings of the* 34*th International Conference on Machine Learning (ICML)* (2017).

10. Anthimopoulos, M. et al. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging. **35**, 1207–1216(2016).

11. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. Science, **361**, 1004–1008(2008).

12. Krizhevsky, A.et al. ImageNet classification with deep convolutional neural networks. Communication of Association for Computing Machinery. **60**, 84–90(2017).

13. He, K. et al. Deep residual learning for image recognition. *In Proc. Conference on Computer Vision and Pattern Recognition (IEEE)*.770–778(2015).

14. Harris, N. C. et al. Quantum transport simulations in a programmable nanophotonic processor. Nat. Photon. **11**, 447–453(2017).

15. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. Nature **588**, 39–47(2020).

16. Zhuang, L. et al. Programmable photonic signal processor chip for radiofrequency applications. Optica **2**, 854–859(2015).

17. Zhang, Q. et al. Artificial neural networks enabled by nanophotonics. Light: Sci. Appl. **8**, 42–56(2019).

18. Feldmann, J. et al. All-optical spiking neurosynaptic networks with self-learning capabilities. Nature **569**, 208–214(2019).

19. Bogaerts, W. et al. Programmable photonic circuits. Nature **586**, 207–216(2020).

20. Wang, J. et al. Integrated photonic quantum technologies. Nat. Photon. **14**, 273–284(2020).

21. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. Nature **577**, 641–647(2020).

22. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nat. Photon. **15**, 367–373(2021).

23. Liu, C. et al. A programmable diffractive deep neural network based on a digital-coding metasurface array. Nat. Electron. **2**, 113–122(2022).

24. Williamson, I. A. D. et al. Reprogrammable electro-optic nonlinear activation functions for optical neural networks. IEEE J. Quantum Electron. **26**, 7700412(2020).

25. Joy, S. R. et al. Spoof plasmon interconnects—communications beyond RC limit. IEEE Trans. Commun. **67**, 599–610(2019).

26. Pendry, J. B. Martin-Moreno, L. Garcia-Vidal, F. J. Mimicking surface plasmons with structured surfaces. Science **305**, 847–848(2004).

27. S. A. Maier, *Plasmonics: Fundamentals and Applications*, Springer, New York 2007.

28. Shen, X. Conformal surface plasmons propagating on ultrathin and flexible films. *Proc. Natl. Acad. Sci. USA.* **110**, 40–45(2013).

29. Zhang, H. C. et al. A plasmonic route for the integrated wireless communication of subdiffraction-limited signals. Light: Sci. Appl. **9**, 113(2020).

30. Gao, X. et al. Crosstalk suppression based on mode mismatch between spoof SPP Transmission line and microstrip. IEEE Trans. Components. Package. Manu. Tech. **9**, 2267–2275(2019).

31. Zhang, H. C. et al. Breaking the challenge of signal integrity using time-domain spoof surface plasmon polaritons. ACS Photonics **2**, 1333–1340(2015).

32. Cui, T. J. Microwave metamaterials. Nat. Sci. Rev. **5**, 134–136(2018).

33. Gao, X. et al. Programmable multifunctional device based on spoof surface plasmon polaritons. IEEE Trans. Antennas Propag. **68**, 1558(2020).

34. Gao, X. et al. Programmable hybrid circuit based on reconfigurable SPP and Spatial waveguide modes. Adv. Mater. Technol. **5**, 1900828(2019).

35. Zhang, J. et al. Integrated spoof plasmonic circuits. Sci. Bulletin **64**, 843–855(2019).

36. Ma, Q. et al. Controllable and programmable nonreciprocity based on detachable digital coding metasurface. Adv. Opt. Mater. **7**, 1901285(2019).

37. Ma, Q. et al. Smart sensing metasurface with self-defined functions in dual polarization. Nanophotonics **9**, 3271–3278(2020).

Figures
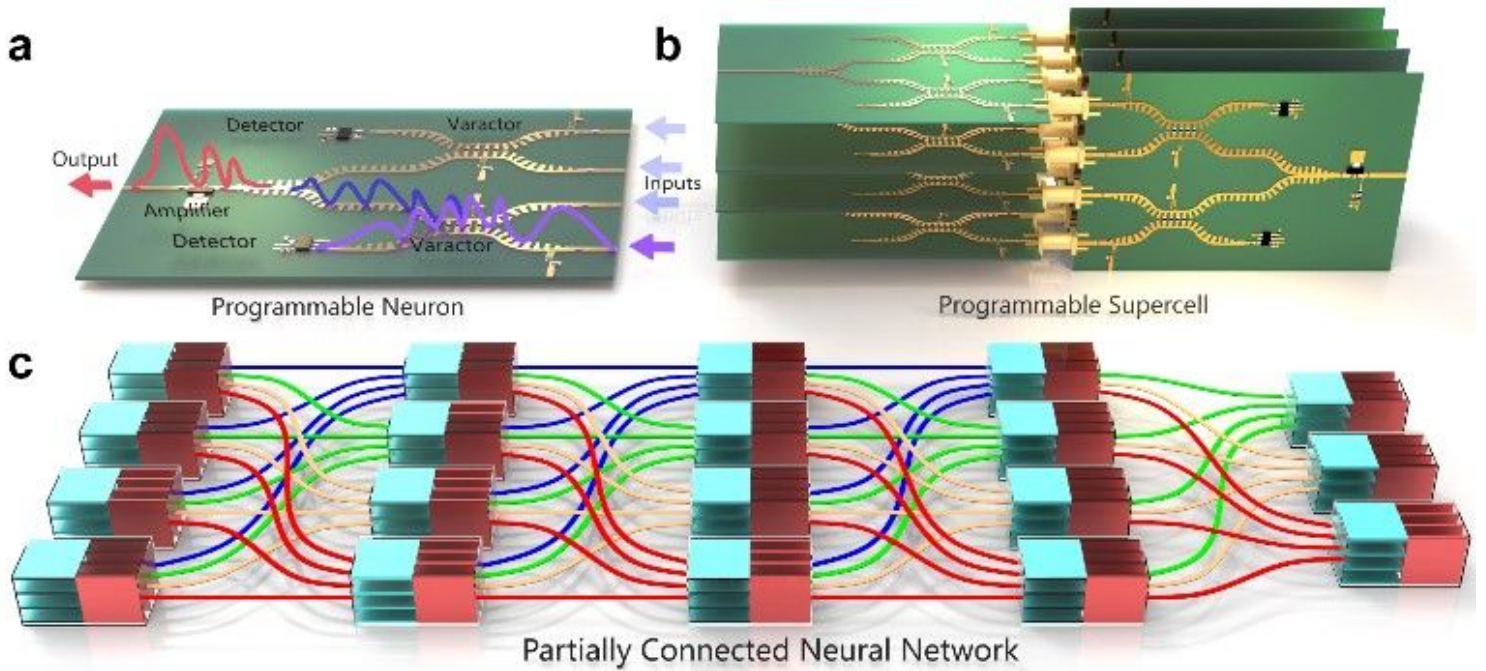
Figure 1

Schematic of the programmable SSPP neuron, supercell, and partially connected SPNN. a, A single programmable SSPP neuron is constructed by integrating the varactors, detectors, and amplifier through the SSPP interconnection with four input ports and one output port. The varactor is controlled by the applied bias voltage to constitute a tunable-ratio power divider, where the input signals are mixed with the customized weight factors. The detectors measure the output signal intensity, which is transformed and fed back to the power amplifier, implementing a programmable activation function. The subfigure illustrates the diagram of an output programmable neuron with detectors and power amplifiers, while the detectors and amplifiers are omitted in the input programmable neurons. **b,** A programmable supercell consisting of eight programmable neurons (four inputs and four outputs) is able to realize arbitrary 4×4 matrix multiplication (See Methods). **c,** A partially connected SPNN with 16 inputs and 12 outputs to carry out the classification task on the MINST dataset. Compared with its fully-connected counterpart, the partially connected SPNN reduces significantly the complexity routing between the neuron layers without performance degradation.
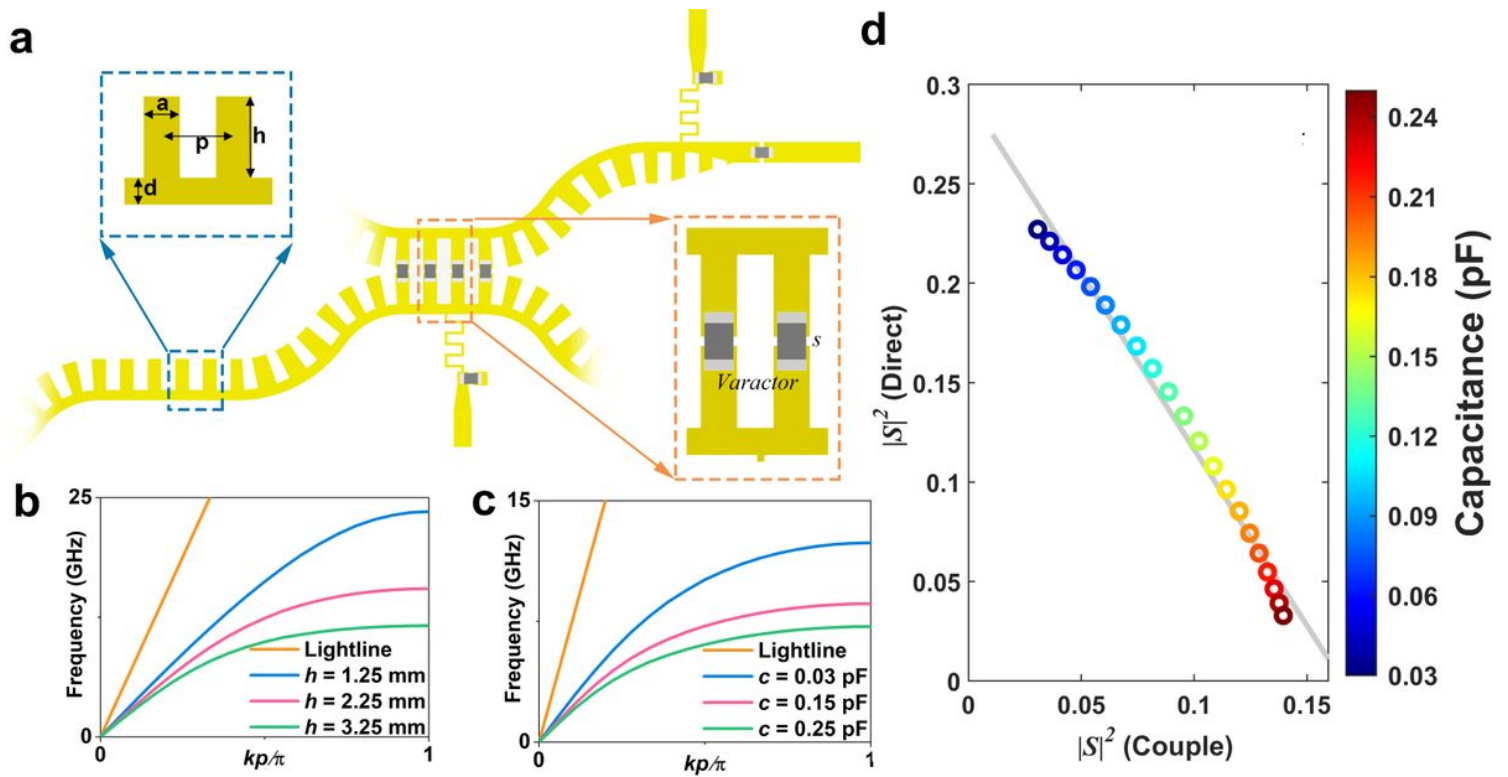
Figure 2

Dispersion behaviors and transmission parameters of the SSPP neuron. a, Unit structure of a single neuron, where the structural parameters are a = 1 mm, d = 0.75mm, p = 2 mm, s = 0.25 mm, and the groove depth and the capacitance of varactor are variable. **b,** The dispersion curves of the unit shown in the left-upper corner in (a) with different groove depths h. **c**, The dispersion curves of the unit shown in the right lower corner in (a) with different capacitances c. **d**, The relation between the input and output signals as the capacitance varies from 0.03 to 0.27 pF, in which the coupling and direct-through coefficients are described in Supplementary Section 1.
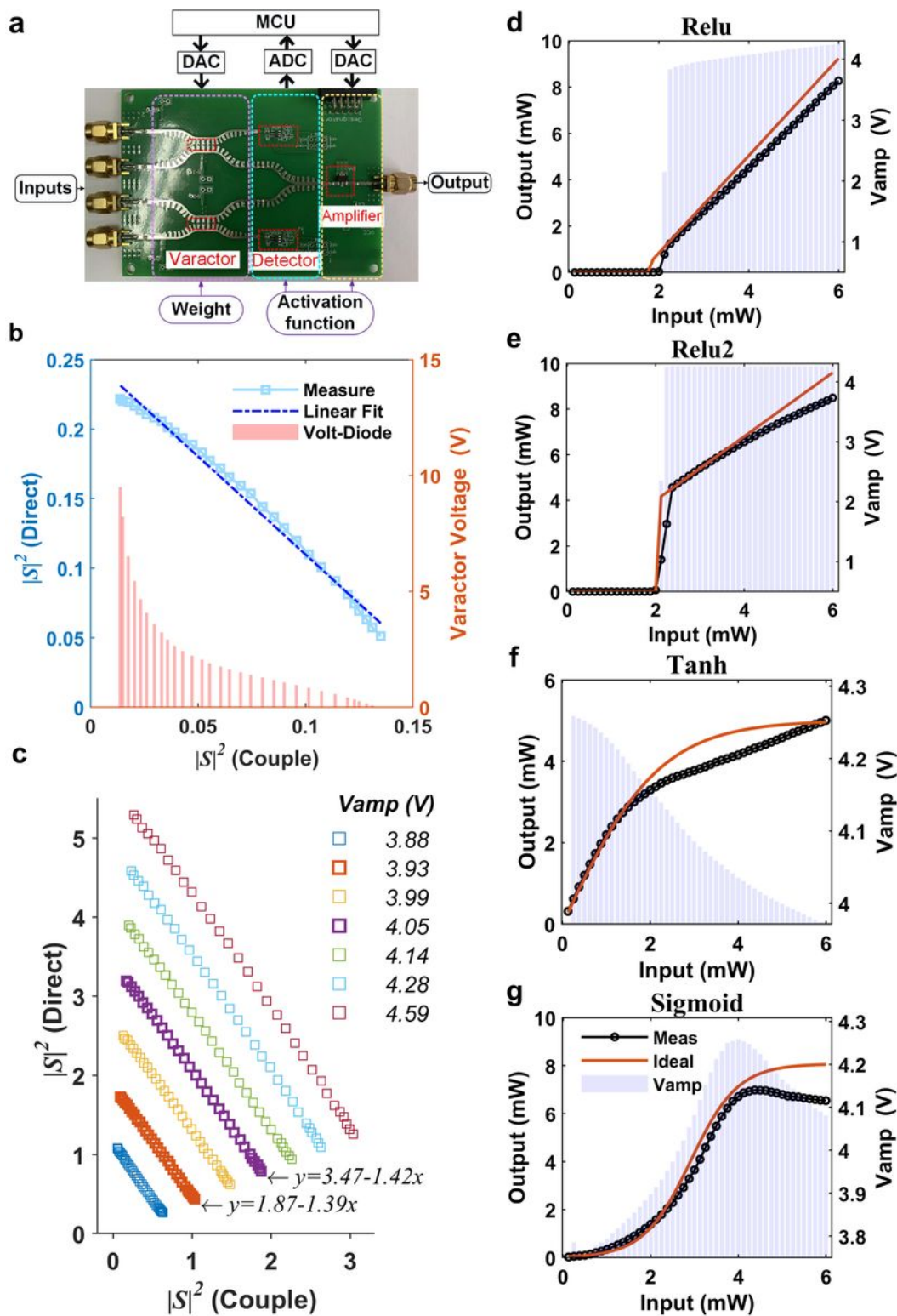
Figure 3

Experimental results of the SSPP output neuron. **a,** Picture of the fabricated sample. **b,** The relationship between the coupling and direct transmissions as the bias voltages of the varactors vary from 0 to 15 V, which denotes that the weight of the neuron is reprogrammable. **c,** The relationship between the coupling and direct transmissions as the bias voltage applied to the amplifier chip varies from 3.88 to 4.59 V. **d, e,**

f, and **g** show four different activation functions, which can be customized by modulating the bias voltage applied to the amplifier, respectively.
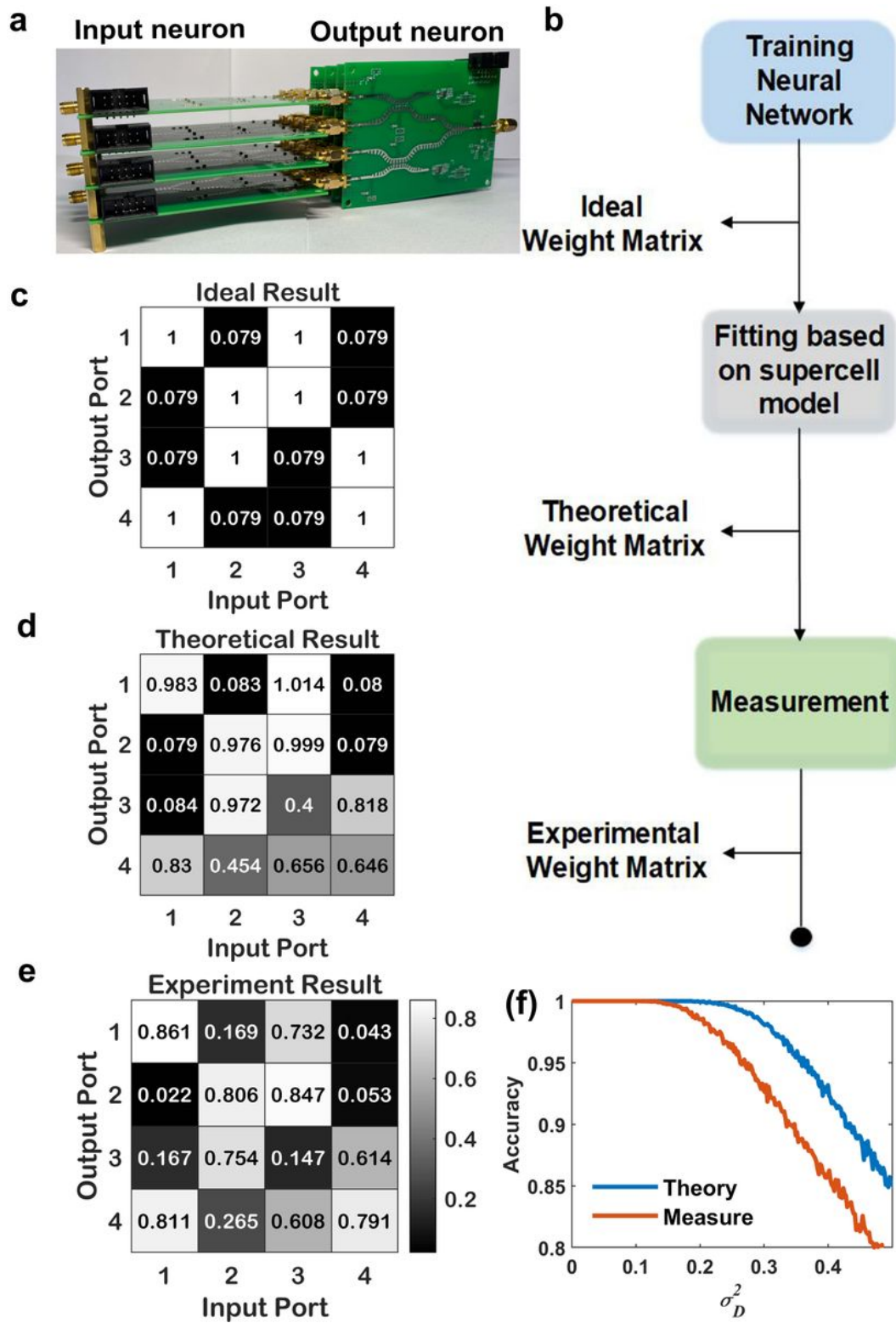

Figure 4

Training performance of the SSPP super-neuron. a, The sample picture of the SSPP super-neuron consists of four input neurons and four output neurons. **b,** The operation process for obtaining the weight

matrices. **c,** The ideal weight matrix between the input and output neurons. (d) The theoretical weight matrix by measuring the super-neuron. **e,** The experimental weight matrix by measuring the super-neuron. **f,** The calculated and measured accuracies of the super-neuron with the Gaussian white noises.
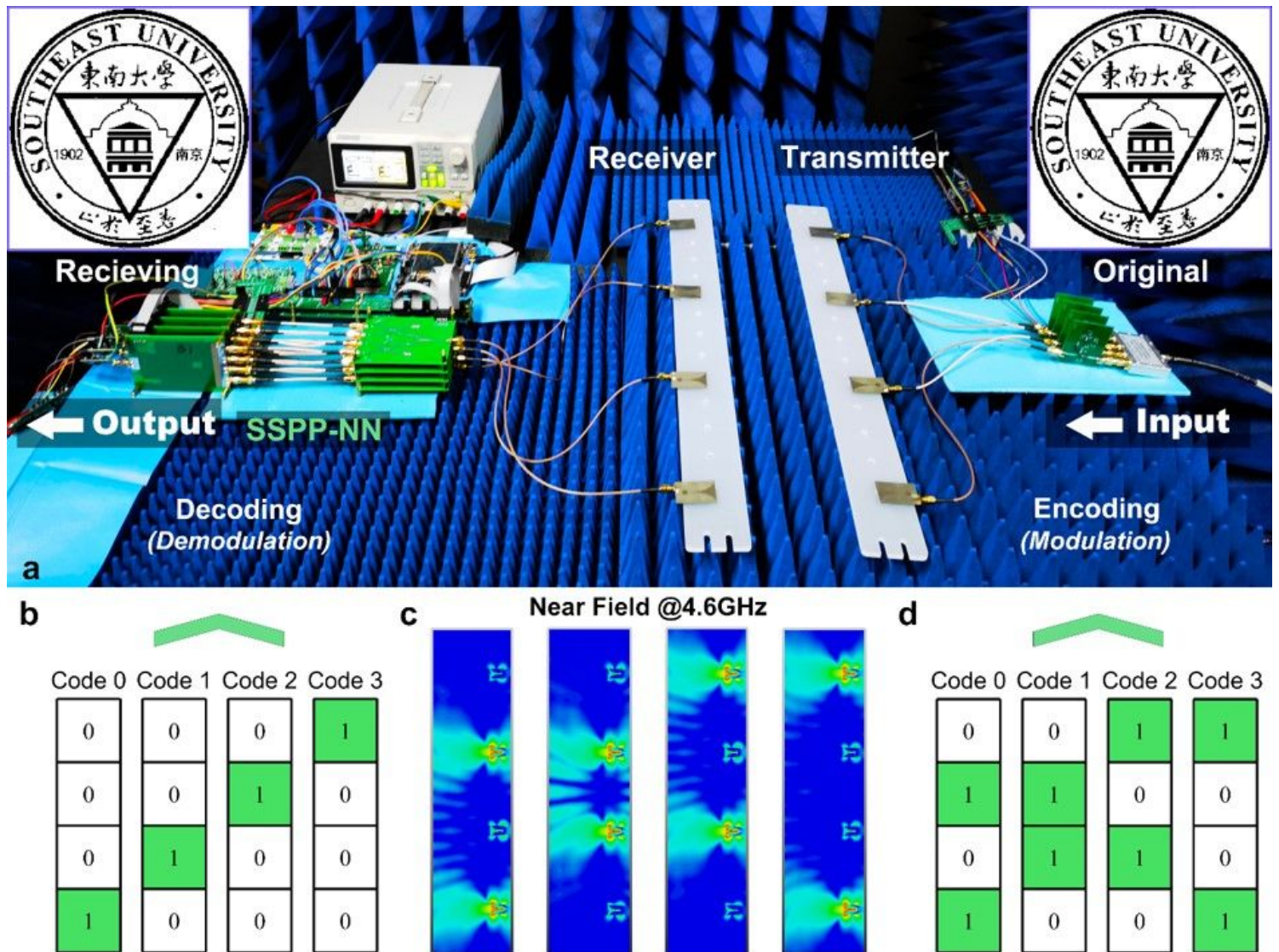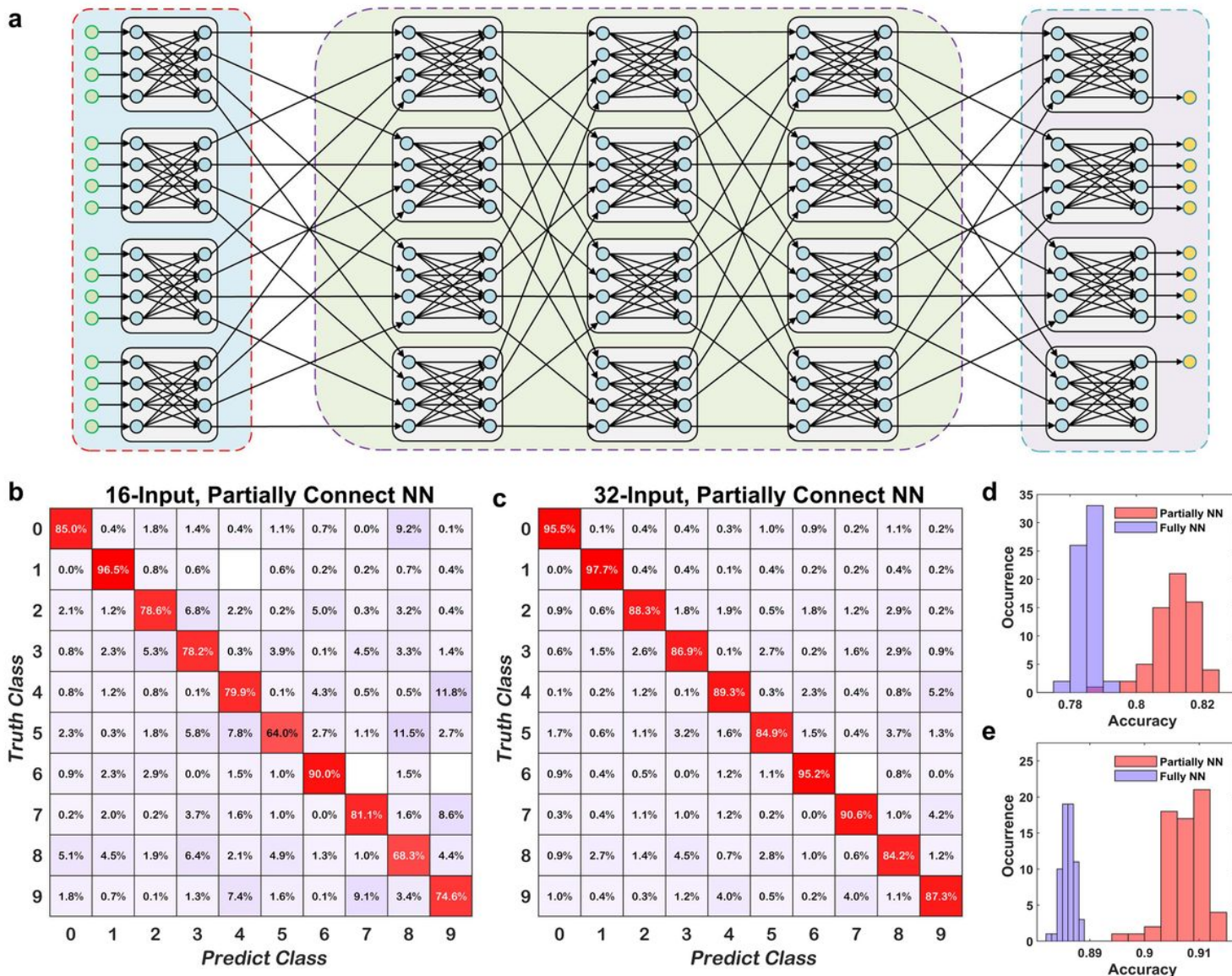

Figure 5

The decoding mechanism in wireless communication system with the SSPP super-neuron. a, Photograph of the communication system, in which a logo picture of Southeast University has been transmitted and received by the SPNN. **b** and **d** are the code mapping between the 2-bit symbols and channel signals in the encoding process and output ports of the SSPP super-neuron. Through the decoding process implemented by the SSPP super-neuron in EM domain, the original code mapping is converted into a simplified injective one. **c,** The near-field patterns between the transmitted antenna array and the received antenna array under four coding circumstances.

Figure 6

Training performance for a partially connected SPNN with the ReLU activation function. a, The schematic diagram of the partially connected SPNN consisting of abundant programmable super-neurons with one input layer, four hidden layers, and one output layer. **b** and **c** are the confusion matrices during training in 16 and 32 inputs, respectively. **d** and **e** are test accuracies of the partially and fully connected ANNs during training in 16 and 32 inputs, respectively. Note that the orange and purple histograms in each subfigure represent the prediction results trained from the partially and fully connected networks, respectively.

Supplementary Files

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

- ProgrammableSurfacePlasmonicNeuralNetworksSupplementaryInformation.docx