Scalability of All-Optical Neural Networks Based on Spatial Light Modulators

Ying Zuo,[1,‡] Yujun Zhao,[1,‡] You-Chiuan Chen,[1] Shengwang Du,[2,*] and Junwei Liu[1,†]

[1] *Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*

[2] *Department of Physics, The University of Texas at Dallas, Richardson, Texas 75080, USA*

(Received 24 January 2021; revised 29 March 2021; accepted 7 April 2021; published 17 May 2021)

Optical implementation of artificial neural networks has been attracting great attention due to its potential in parallel computation at speed of light. Although all-optical deep neural networks (AODNNs) with a few neurons have been experimentally demonstrated with acceptable errors recently, the feasibility of large-scale AODNNs remains unknown because error might accumulate inevitably with increasing number of neurons and connections. Here, we demonstrate a scalable AODNN with programmable linear operations and tunable nonlinear activation functions. We verify its scalability by measuring and analyzing errors propagating from a single neuron to the entire network. The feasibility of AODNNs is further confirmed by recognizing handwritten digits and fashions, respectively.

DOI: 10.1103/PhysRevApplied.15.054034

I. INTRODUCTION

In the last decades, artificial neural networks have grown into one of the most disruptive technologies [1–3] and found success in both practical applications, such as computer vision, speech recognition, and natural language processing [4–6], and fundamental studies such as particle physics, condensed-matter physics, and material science [7–17]. The power of an artificial neural network comes from its large amount of artificial neurons and their intensive interconnections, which require large memory size, computational complexity, and energy consumption [18]. Due to the light wave nature, all-optical deep neural networks (AODNNs) hold great promise in high-speed parallel computation with low energy consumption [19–26]. Research in neuromorphic computing has flourished [25,27,28] in both dense neural networks and convolutional neural networks [29,30]. AODNNs have not been implemented experimentally until recently with the breakthroughs in realizing optical nonlinear activation functions using phase-change materials [31], electromagnetically induced transparency (EIT) atomic medium [32], and saturated absorption [33], which explicitly remove all the principle restrictions of constructing a general-purpose AODNN. However, any of these demonstrated AODNNs has no more than 22 neurons and the scalability remains a question.

Different from a universal digital electronic implementation of artificial neural networks, an optical neural network is usually designed and built for a specific given task [34–36], and one has to rebuild a completely new optical neural network even for a slightly different task, which shares a similar structure. In addition, the error from the imperfection of optical signals and components may accumulate exponentially with the number of neurons to completely ruin the performance in a large optical neural network even though in general large-size neural networks have large tolerance for random local error from an individual neuron [37].

In this work, we demonstrate a scalable AODNN based on free-space Fourier optics and EIT nonlinear activation functions. The programmable linear transformations in our AODNNs are realized using spatial light modulators (SLMs) [38] and optical lenses. We experimentally show that local random errors can be reduced by increasing the area of SLM (S) per neuron. Moreover, the total error in the linear transformation accumulates only with the deviation proportional to the square root of the number of neurons (M). In addition, the nonlinear activation functions realized with EIT in cold atoms are spatially separated, thus their errors are independent and accumulates similarly to the linear transformation (with the deviation proportional to the square root of their number). Therefore, this AODNN can be scaled up to large size since the effect of this kind of slowly accumulated errors can be easily compensated by increasing the number of hidden neurons [37]. We confirm the feasibility and programmability by experimentally constructing a fully connected AODNN with 174

*dusw@utdallas.edu

†liuj@ust.hk

‡These authors contributed equally to this work.

optical neurons, and apply it to recognize handwritten digits and fashions, with classification rates of 81.8% and 71.3%, respectively. Furthermore, we clearly show its scalability using accurate numerical simulations, where the classification rates indeed increase with the number of optical neurons even with large random error.

II. OPTICAL NEURONS AND INTERLAYER CONNECTIONS.

Figure 1 shows the schematics of optical neurons and their interconnections between two adjacent layers. As the building block of an artificial neural network, neurons receive input signals u_j^{in}, conduct a linear transformation $z_i = \sum_j W_{ij} u_j^{\text{in}}$, pass the results to the nonlinear activation function $f_i(\bullet)$, and produce the output signals $u_i^{\text{out}} = f_i(z_i)$, as shown in Fig. 1(a). In our AODNN, signals are encoded in the power of light, and computations are conducted through light propagation, diffraction, and interference. In detail, we perform a linear transformation using a SLM and a Fourier lens, as shown in Fig. 1(b). The input light beams ($u_1^{\text{in}} \sim u_M^{\text{in}}$) are incident on different parts of the SLM, which is placed at the back focal plane of the lens. Through controlling the multiple phase gradings in a certain area of the SLM, the input beam array u_j^{in} are diffracted to different directions with powers $W_{ij} u_j^{\text{in}}$. Then, these weighted beams from different neurons are focused onto the same spot on the front focal plane of the lens and complete the linear summation $z_i = \sum_j W_{ij} u_j^{\text{in}}$. It is worth noting that the computation time here is independent of the number of neurons and interlayer connections but determined by the propagation time of light from the SLM to the front focal plane of the lens. The wave nature of light makes it possible to perform linear transformation with massive

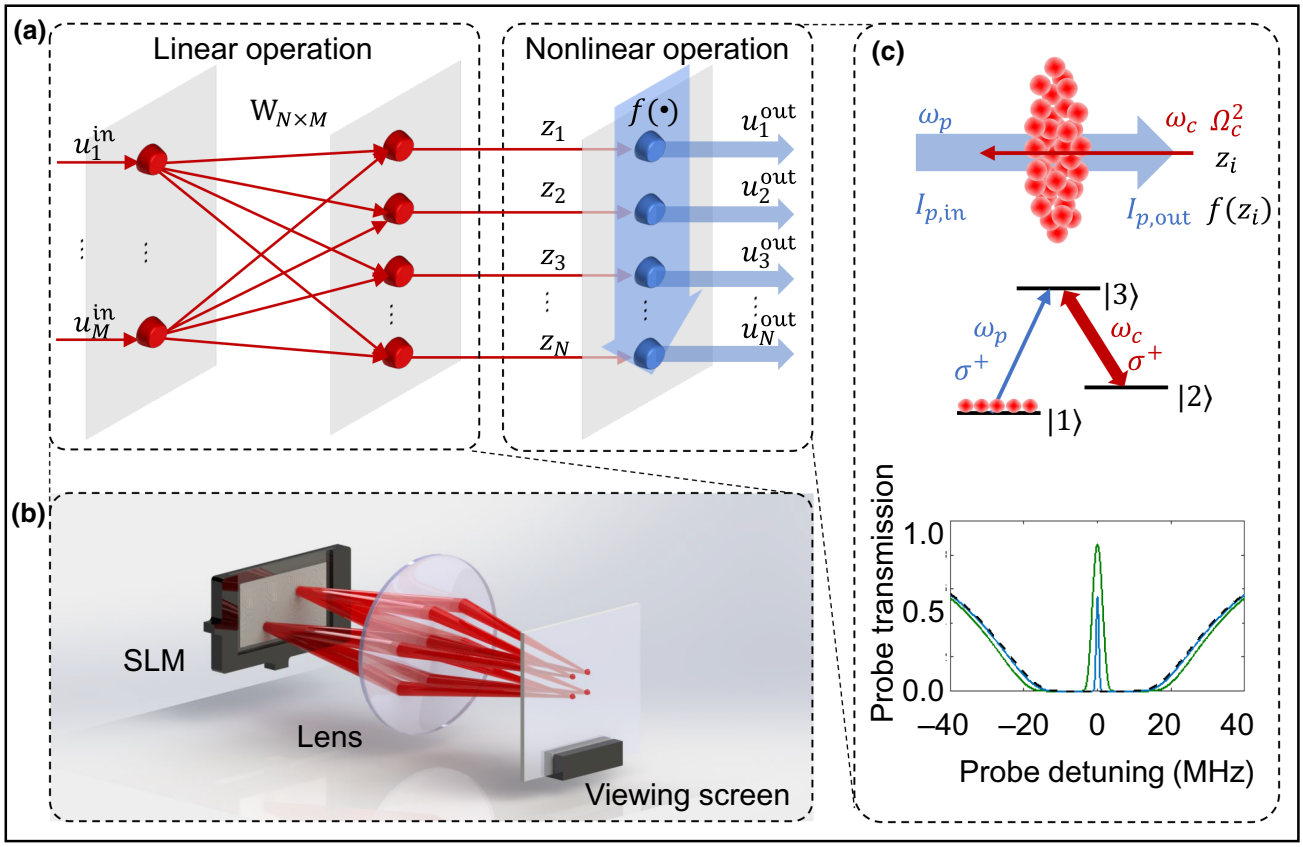

FIG. 1. Optical neurons and connections. (a) Schematics of optical neurons and connections between two adjacent layers in an AODNN. The M nodes u_j^{in} form the input layer. $\mathbf{W_{N\times M}}$ is the $N \times M$ linear linear matrix connection. $f_i(z_i)$ are N outputs after the nonlinear activation functions $f_i(\bullet)$. (b) Optical implementation of linear matrix operation $\mathbf{W_{N\times M}}$. A SLM, placed on the back focal plane of an optical lens, splits each input optical nodes into multiple directions. The lens performs Fourier transform and sum the light beams along the same directions onto a spot on its front focal plane. (c) EIT nonlinear optical activation function with a three-level atomic medium (top panel) and EIT transmission spectrum of the probe beam (bottom panel). The transmission of the probe beam power $I_{p,\text{in}}$ is nonlinearly controlled by the power I_c of the control light. The relevant ^{85}Rb atomic energy levels are $|1\rangle = |5^2S_{1/2}, F = 2\rangle$, $|2\rangle = |5^2S_{1/2}, F = 3\rangle$, and $|3\rangle = |5^2P_{1/2}, F = 3\rangle$. The black dashed line in the bottom panel is obtained without the control beam and shows maximum absorption on resonance at zero probe detuning. The solid lines (blue and green) are two typical EIT spectra with different control light power.

parallelism at the speed of light. Moreover, different from many other systems, the layer structure, including the number of neurons and the weight matrix, in our system are fully programmable.

The optical nonlinear activation functions are realized using EIT [39,40] effect in cold ^{85}Rb atoms prepared in a two-dimensional (2D) magneto-optical trap (MOT) [41,42]. As shown in the top panel of Fig. 1(c), a pair of counterpropagating probe (ω_p) and control (ω_c) beams are shined on the atomic medium. The relevant ^{85}Rb atomic energy levels are labeled as $|1\rangle$, $|2\rangle$, and $|3\rangle$, and the atoms are in the state $|1\rangle$. The circularly (σ^+) polarized probe laser is on resonance to the transition $|1\rangle \rightarrow |3\rangle$, and the circularly ($\sigma^+$) polarized control laser is on resonance to the transition $|2\rangle \rightarrow |3\rangle$. Without the control light, the atomic medium is opaque to the probe beam due to its on-resonance absorption. In the presence of the control beam, the atomic medium becomes transparent and the probe transmission depends on the power of the control light. The bottom panel of Fig. 1(c) shows typical probe transmission spectrums with different control light powers. The nonlinear dependency of the output probe power (I_p^{out}) on the control laser power (I_c) is described as $I_p^{\text{out}} = f(I_c) = I_p^{\text{in}} e^{-\text{OD}(4\gamma_{12}\gamma_{13}/\Omega_c^2+4\gamma_{12}\gamma_{13})}$, where I_p^{in} is the input probe beam power. Ω_c is the Rabi frequency of the control beam and Ω_c^2 is proportional to control beam intensity I_c. Here, $\gamma_{13} = 2\pi \times 3$ MHz is fixed and determined by the spontaneous emission of the excited state $|3\rangle$. The ground-state dephasing rate γ_{12} can be engineered by applying external magnetic field. OD is the atomic optical depth on the probe transition and can be varied by changing the atomic density. In our optical neuron, the control light works as the input signal to its nonlinear activation function and the probe light output is the output signal of the activation function. The EIT optical nonlinear activation function can be tuned by manipulating the cold atom parameters OD and γ_{12}, and they are also independent for different neurons. Under the experimental condition that OD ≈ 2 and $\gamma_{12} \approx 5\gamma_{13}$, the activation function shows strong nonlinearity with coupling power to be less than 1 mW with the diameter of beam to be 100 μm.

III. ERROR AND SCALABILITY

Clearly, the scalability of our AODNNs is determined by the accuracy of the linear interlayer connections $\mathbf{z}_N = \mathbf{W}_{N\times M}\mathbf{u}_M^{\text{in}}$, where $\mathbf{z}_N$ ($\mathbf{u}_M^{\text{in}}$) is a N- (M-) dimensional vector, and $\mathbf{W}_{N\times M}$ is an $N \times M$ matrix. In our optical implementation, such a linear transformation is divided into two steps: (1) weighted multiplication by the SLM: for any given $j \in [1, M]$ and $i \in [1, N]$, $y_i^j = W_{ij}u_j^{\text{in}}$; and (2) summation by the lens: $z_i = \sum_{j=1}^{M} y_i^j$. M and N are the number of input and output nodes. As illustrated in Fig. 1(b), the SLM is divided into M sections and the beam on each section is split into N directions. Figure 2(a) displays typical optical intensity distribution patterns of $M = 144$ input nodes on the SLM and $N = 64$ output nodes split from one of the input nodes. Obviously, the M input sections ($\mathbf{u}^{\text{in}}$) on the SLM are independent. For a given input node u_j^{in}, its splitting into N weighted outputs, i.e., $y_i^j = W_{ij}u_j^{\text{in}}$ are handled by the same section area S of SLM and inevitably entangled with each other.

We use the weighted Gerchberg-Saxton (GSW) algorithm [43–45] to obtain an appropriate phase pattern applied on SLM for achieving target weights W_{ij}. At each iteration of the GSW algorithm, we measure powers of output spots and compute the root-mean-square-error (RMSE) between measured weights and target weights, which is taken as feedback for the next iteration (more details are shown in Appendix A). A typical example of the iteration process is shown in Fig. 2(b), where the RMSE rapidly decreases within 20 iteration steps and usually converges in less than 50 steps. To make our test experiment fair, we randomly choose the value of weights between 0 and 1. The test result for $N = 400$ and $S = 320 \times 180$ is shown in Fig. 2(c). Overall, the accuracy is very high with a small error, and the error is slightly larger for target weights around zero because it is difficult to reduce the light power to zero due to the diffraction from the adjacent beams. As shown in the insert of Fig. 2(c), the distribution of absolute error is close to a normal distribution with a single symmetric peak centered at zero with a standard deviation of 0.019, which indicates the random nature of the error and quantitatively show the high accuracy of the linear transformation. We find that the RMSE between measured weights and target weights depends only on $\sqrt{N/S}$, as shown in Fig. 2(d). For a small N, the RMSE is linearly proportional to $\sqrt{N}$ for all different area sizes S of the input node and the slope of the fitted line is proportional to $1/\sqrt{S}$ (see Sec. I within the Supplemental Material [46] for more details). Such a dependency of error on N and S is governed by the nature of Fourier transformation (see Sec. II within the Supplemental Material [46] for the theoretical analysis). As N becomes very large, the nearest diffracted spots cannot be well separated, which increases the error significantly. The threshold is indicted by the gray lines in the figure and it clearly increases with S because the area of each diffracted spot decreases with S.

We then analyze the error accumulated in step (2) of light power summation by comparing the measured output $\tilde{z}_i$ to the target values $z_i = \sum_j w_{ij}u_j$. As shown in Fig. 2(e), for a given S, the RMSE(M, N, S) for different N is linearly proportional to $\sqrt{M}$. We also find that for a given N, the RMSE(M, N, S) for different S is linearly proportional to $\sqrt{M}$ as shown in Fig. 2(f) (see Sec. I within the Supplemental Material [46] for more details). Such a $\sqrt{M}$ dependency of RMSE(M, N, S) confirms that the errors

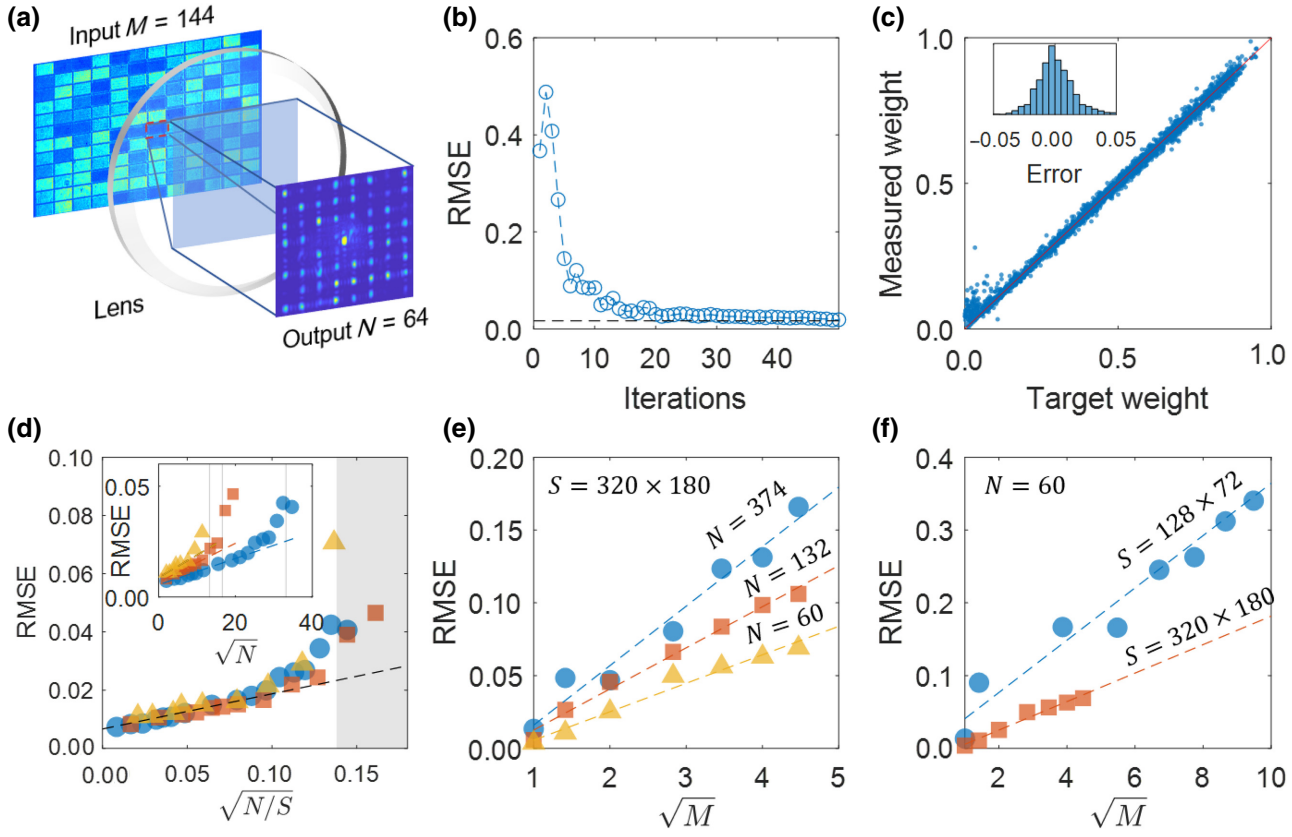

FIG. 2. Scalability of linear operation $\mathbf{W}_{N\times M}$. (a) Typical images of input nodes vector $\mathbf{u_M}^{\text{in}}$ and the weighted output $W_{ij}\,u_j^{\text{in}}$ from an individual input node u_j^{in}. For the purpose of illustration, we take here $N = 64$ and $M = 144$. (b) Error during the GSW feedback iteration process for configuring the weights W_{ij}. The data are taken with $S = 320 \times 180$ and $N = 400$. (c) The measured weights versus the target weights. The inset histogram shows the error distribution. (d) The dependency of error on $\sqrt{N/S}$. The yellow triangles, red squares, and blue circles are the experimental RMSE data of $S = 128 \times 72$, $S = 160 \times 90$ and $S = 320 \times 180$ accordingly. The dashed lines are the fitting curves of each case. (c),(f) The dependency of error on $\sqrt{M}$ with fixed S and N, respectively.

from different input nodes are random, independent, and uncorrelated, and the total error accumulates only in the stochastic way. Combining with the previous analysis of error dependency on N and S, we conclude that the error of optical linear summation increases linearly with $\sqrt{MN/S}$.

It is worth noting that, in our error analysis, N actually refers to the number of beams diffracted from a single-node area of SLM (one neuron in the sense of neural network), whose area is labeled by S, and it is equal to the number of neurons in the next adjacent layer for a fully connect neural network. Then, the actual meaning of MN should be the number of links between two adjacent layers. Therefore, we can reduce the error of optical linear summation if building a convolutional all-optical neural network. Another thing to notice is that the linear combination error mentioned is absolute error, which means the relative error decreases with $\sqrt{N/SM}$.

With all errors confirmed to remain local and random, the scalability of our AODNNs is mainly limited by the capacity of linear matrix operation: number of optical neurons and their interconnections between two adjacent layers. For a given pixel size ($d \times d$) and number of pixels (K) of the SLM, the numerical aperture (NA) of the Fourier lens, as well as the light wavelength λ, our analysis shows the maximum connections, or the linear capacity of a single SLM-lens layer, is determined by $C_L = (M \times N)_{\max} = (\text{NA}^2\pi K d^2/4\lambda^2)$ (more details in Appendix B). For the SLM (HOLOEYE, PLUTO-2) used in this work, we have $d = 8\mu$m, $K = 1920 \times 1080$. Together with NA $= 0.05$ of the lens and the light wavelength $\lambda = 795$ nm, our single SLM-lens layer can accommodate up to $C_L = 412,287$ connections. In experiment, due to the restricted size of the camera (Hamamatsu C11440-22CU), the maximum connection number we test in Fig. 2(d) is 46 656. Our system still has potential to be scaled up.

IV. AODNN APPLICATION EXAMPLES.

We confirm the feasibility and programmability of our AODNNs by building a fully connected two-layer all-optical neural network to recognize both handwritten digits and fashion images. The network structure and its optical

layout are shown in Figs. 3(a) and 3(b). There are 144, 20, and 10 optical neurons in the input, hidden, and output layer, respectively. A collimated control laser beam (marked with red color) with horizontal linear polarization, generated from a fiber output through lens L1 is shone on SLM1 that is placed on the back focal plane of lens L2. The SLM1 is divided into 144 sections and reflects the control light to produce 144 input nodes with controllable weights through the programmed phase gratings on each section (see Sec. III and Fig. S1–S3 within the Supplemental Material [46]). An aperture stop is placed on the front focal plane of L2 to filter away unwanted high-order diffraction modes. Then these 144 spatial nodes are imaged through the lens L3 on the SLM2. The combination of the SLM2 and the lens L4 performs the first linear matrix operation $\mathbf{W_1}$. The 20 outputs of this linear operation are then imaged on the cold atoms in MOT through a telescopic setup of the lenses L5 and L6. A quarter-wave plate (QWP1) converts the horizontal linear polarization into σ^+ circular polarization to meet the EIT requirement. The control-beam transverse-mode area inside the MOT is about 400 μm^2. A probe laser beam (marked with blue color, σ^+ circularly polarized after QWP2) from the fiber output is collimated by the lens L7 with a beam diameter 8 mm and is shone to the MOT with the opposite direction to the control beams (see Sec. IV and Fig. S4 within the Supplemental Material [46]). The nonlinear functions of different nodes are not exactly the same since OD, γ_{12}, and probe-beam intensity are spatially varied. The probe beam, after passing through the atomic medium spatially dressed by the control beams, contains the output spatial patterns of the 20 hidden neurons. Passing through QWP1, the probe-beam node becomes vertically polarized and is reflected by the PBS. The 20 hidden neuron outputs are then imaged on the SLM3 through the lenses L6 and L8. The combination of SLM3 and L9 performs the second linear operation $\mathbf{W_2}$ and generates ten output nodes recorded by a camera.

We first test the ability of the AODNN for handwritten digit recognition. The initial training is done with a computer-based neural network with the same network structure and nonlinear activation functions (see Sec. XIII within the Supplemental Material [46] for details of training). We train the classifier using the Modified National Institute of Standards and Technology (MNIST)

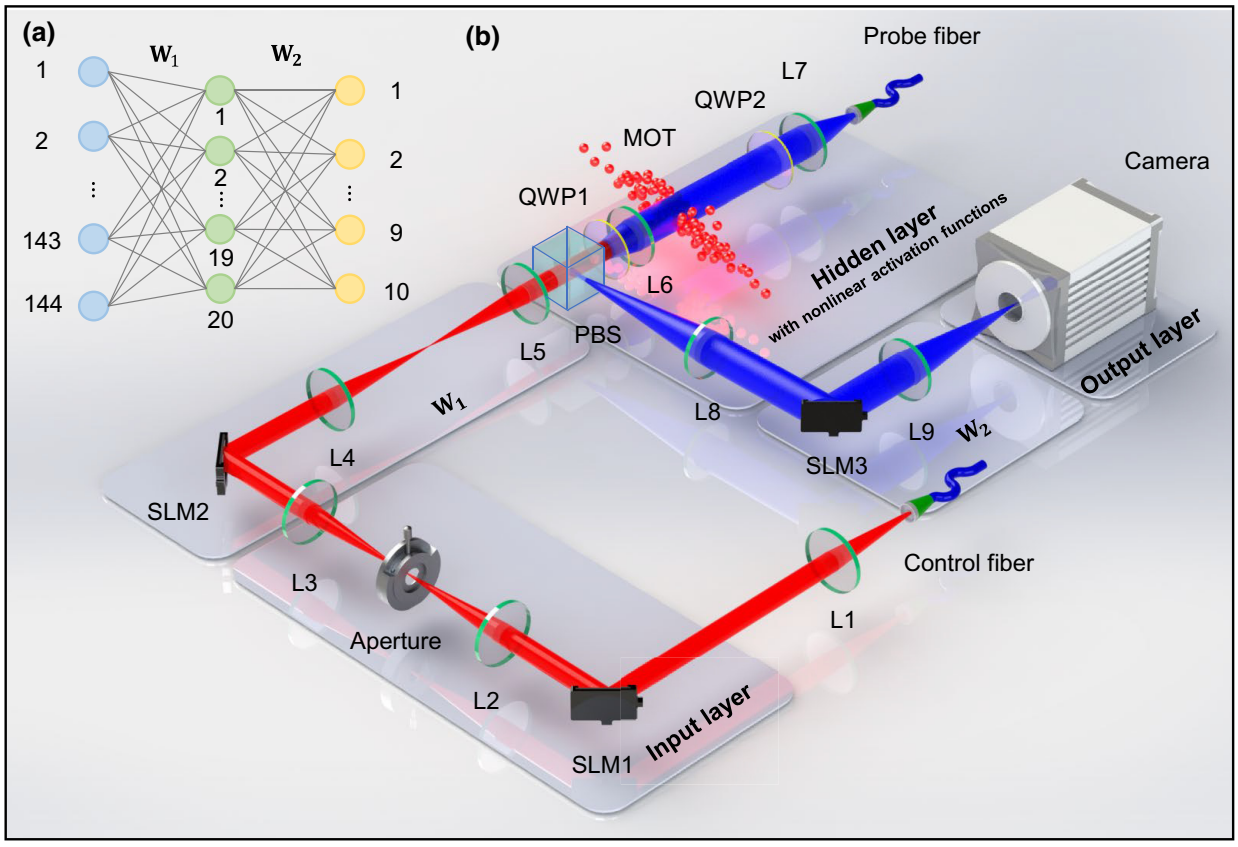

FIG. 3. A two-layer AODNN with 144 input neurons, 20 hidden neurons, and 10 output neurons. (a) The corresponding artificial neural-network architecture. (b) The optical layout of the AODNN. Spatial light modulators: SLM1 (HOLOEY LETO), SLM2(HOLOEYE PLUTO-2), and SLM3(HOLOEY GEAE-2). Camera, Hamamatsu C11440-22CU; PBS, polarization beam splitter. Lenses, L1 (f = 100 mm), L2 (f = 200 mm), L3 (f = 250 mm) L4 (f = 350 mm), L5 (f = 350 mm), L6 (f = 50 mm), L7 (f = 50 mm), L8 (f = 450 mm), and L9 (f = 450 mm).

handwritten digit database [47] [Fig. 4(a), compressed into images with 12×12 resolution] and obtained the optimal parameters of the two linear matrix operations $\mathbf{W_1}$ and $\mathbf{W_2}$(see Supplemental Material [46] for MatrixElement.mat). We achieve a classification rate of 81.8%, and the corresponding normalized confusion matrix is shown in Fig. 4(b), which is very close to the computer-trained normalized confusion matrix in Fig. 4(c). The AODNN hardware is programable. As a second example, we retrain the network with the fashion images [48] [Fig. 4(d)] and configure the AODNN with completely different parameters (see Supplemental Material [46] for MatrixElement.mat). We obtain a classification rate of 71.3%. The normalized confusion matrix of the AODNN [Fig. 4(e)] also agrees well with that of the computer-trained neural network [Fig. 4(f)]. To further demonstrate the ability of our AODNN, we use the computer to simulate the results with more hidden neurons due to the limited physical resources available in our lab. We first simulate the dependency of classification rate on the local errors that are added on each neuron randomly and independently. We perform 1000 different simulations with random relative local errors generated from Gaussian distribution $\mathcal{N}(0, \sigma^2)$, where σ is the standard deviation of relative local error. As shown in Fig. 5(a), the classification rate indeed drops with local error as expected. However, the drop is relatively slow, and encouragingly the classification rate still can achieve 71.5% for the large relative error 0.25. We choose the random relative error of standard deviation $\sigma = 0.15$ to simulate the dependency of classification rate on the number of hidden neurons. As shown in Fig. 5(b), the classification rate increases with the number of

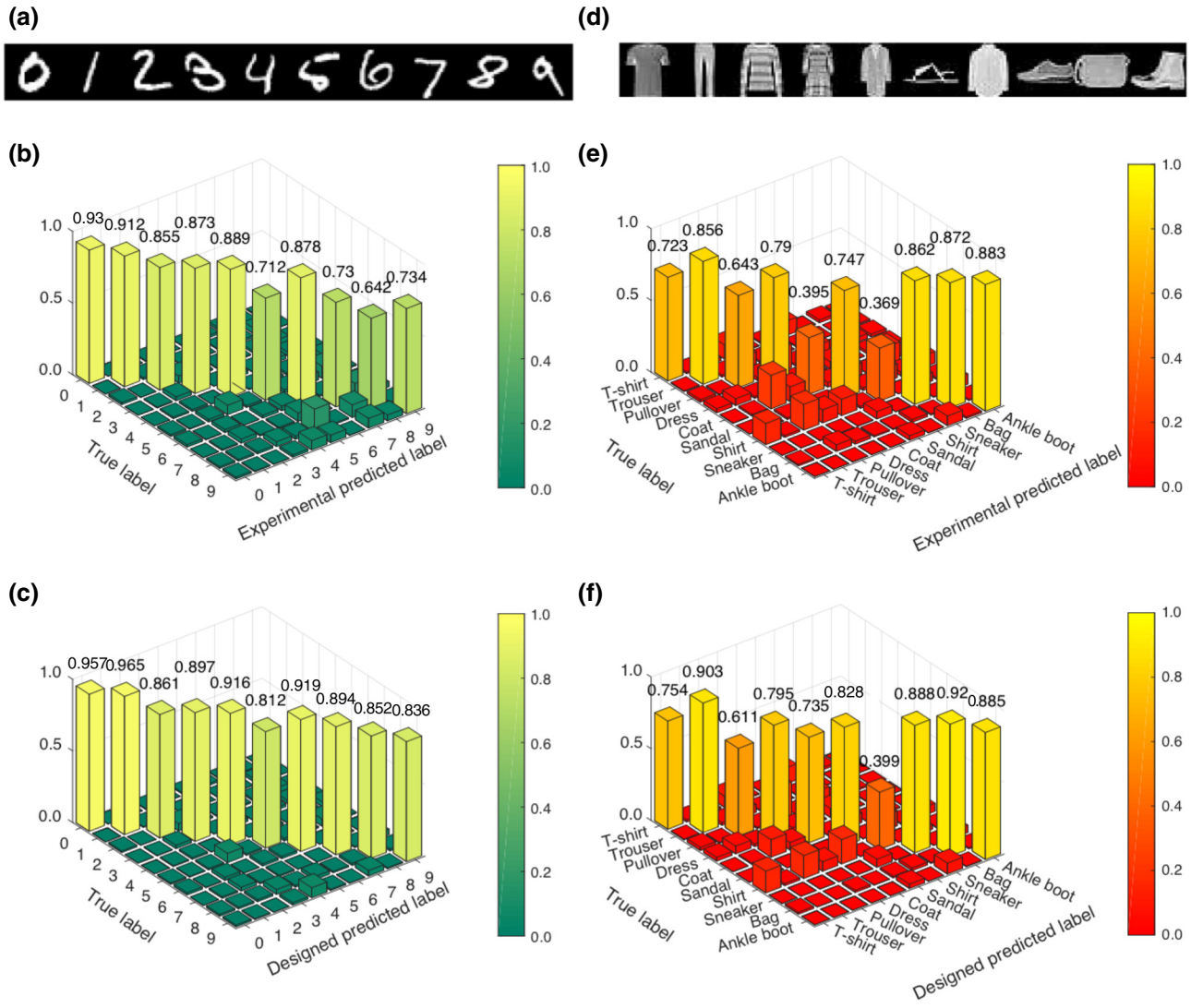

FIG. 4. Results of handwritten digit and fashion classifications. (a),(d) Examples of the test set of ten handwritten digits and fashion patterns. (b),(e) The measured normalized confusion matrixes of the AODNN. (c),(f) The normalized confusion matrixes of a computer-based neural network with the same structure as the AODNN. Both test sets have 8000 patterns.

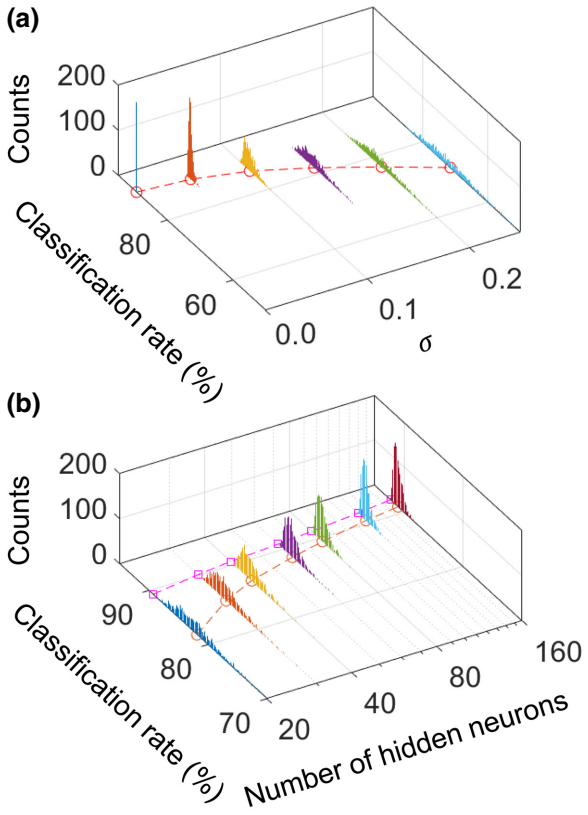

FIG. 5. Classification rate of a two-layer AODNN simulation as a function of (a) standard deviation of random relative local errors and (b) the number of hidden neurons. (a) Numerical simulation of a two-layer AODNN by fixing 144 input neurons, 20 hidden neurons, and 10 output neurons. Random relative errors sampled from Gaussian distribution $\mathcal{N}(0,\sigma^2)$ are added to the weight matrix elements with σ from 0 to 0.25. (b) Numerical simulation of two-layer AODNNs by fixing 144 input neurons, ten output neurons and varying the number of hidden neurons. Random relative errors sampled from Gaussian distribution $\mathcal{N}(0,\sigma^2)$ is added to the weight matrix elements with $\sigma = 0.15$. On both (a) and (b), we perform a different 1000 simulations with each combination of σ and the number of hidden neurons. The histogram shows the classification rate distribution of 1000 simulations. The red circles show the average classification rate of 1000 simulations. The squares present the classification rate of simulation without error. The two-layer AODNNs are trained with the MNIST database.

hidden neurons, and the variance of the classification rate also becomes smaller for a large number of hidden neurons. It means that the effect of local random error becomes smaller and smaller in a larger-size AODNN. Due to the high consistency between our experiments and numerical simulations as proved on the recognition on both handwritten digit and fashion images above, we can expect similar behaviors in experiments when enough physical source is available. These results demonstrate that the drop of the classification rate induced by the local error can be easily compensated by increasing the number of hidden neurons. In addition, we also numerically demonstrate that the usage of EIT nonlinear activation function can indeed improve the classification successful rate. See Sec. IX and Fig. S7 within the Supplemental Material [46] for details.

V. CONCLUSION

Through a systematical error measurement and analysis, we show that the AODNN based on SLM, Fourier optics, and EIT nonlinear activation functions is scalable. We find that the errors remain local, random, and independent, ($\propto \sqrt{N/S}$) and the error accumulation follows only the square-root law of the network dimensions ($\propto \sqrt{MN/S}$), and hence the effect of these errors can be easily compensated by increasing the number of hidden neurons [37]. We confirm the feasibility and programmability of AODNN with handwritten digit and fashion image recognitions and obtain comparable classification rates to computer-based neural networks with the same structure. Although our experimental demonstrations take the example of a two-layer AODNN architecture with only one hidden layer limited by our physical resource in lab, it can be extended into a much deeper AODNN. In the next hidden layer, we can prepare atoms in state $|2\rangle$ in a second MOT, where the probe beams from the previous hidden layer serve as "control" signals for the EIT nonlinearity and control beams become "probe" outputs. The nonlinear activation functions based on MOTs with atoms prepared in state $|1\rangle$ and $|2\rangle$ can be reused if capacity of MOT allows. By carefully designing the EIT parameters (such as atom number and ground-state dephasing rates), a functional deeper AODNN with multiple hidden layers is achievable. For an AODNN, we can also insert optical repeaters (such as a laser power amplifier) to compensate the passive losses in a very deep network. The transverse size can be simply extended with more SLMs and lenses. Our results suggest that a larger-scale AODNN will be promising for light-speed optical computation and could be a powerful AI hardware.

ACKNOWLEDGMENTS

Y.C.C. acknowledges the support from the Undergraduate Research Opportunities Program at the Hong Kong University of Science and Technology. Y.Z. and J.L. acknowledge the support from the Hong Kong Research Grants Council (Grants No. ECS26302118, No. 16305019, No. 16306220, and No. N_HKUST626/18).

APPENDIX A: GERCHBERG-SAXTON ALGORITHM

The weighted Gerchberg-Saxton algorithm [49] is a widely used algorithm. We develop an adaptive GSW iteration program to optimize the phase patterns of a SLM to obtain target weights in a linear transformation. To do

so, we insert a flip mirror to the AODNN to reflect the light beams and image them into a camera. The difference between the measured intensity pattern and the target is fed back to adjust the phase pattern of the SLM until desired weights with acceptable errors are achieved. See Sec. V and Fig. S5 within the Supplemental Material [46] for details of realization of GSW algorithm.

APPENDIX B: LINEAR CAPACITY OF A SINGLE SLM-LENS LAYER

We use a SLM and a Fourier lens to perform a linear matrix operation. There are total K pixels in the SLM and the area of each pixel is d^2. With total M input optical neurons, each node occupies a rectangular area of Kd^2/M. On the front focal plane of the Fourier lens with a focal length f, the corresponding Fourier transformed spot area is $4Mf^2\lambda^2/(Kd^2)$ (see Sec. VI within the Supplemental Material [46] for calculation of capacity). With the lens diameter D, the maximum number of nodes on the Fourier plane is $N_{\max} = (K\pi D^2 d^2/4Mf^2\lambda^2)$. Then the linear capacity is estimated as $C_L = M \times N_{\max} = (\mathrm{NA}^2\pi Kd^2/4\lambda^2)$, where $\mathrm{NA} = D/(2f)$ is the numerical aperture of the lens.

[1] M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science **349**, 255 (2015).

[2] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature **559**, 547 (2018).

[3] E. Alpaydin, *Introduction to Machine Learning* (MIT press, Cambridge, MA, 2020).

[4] D. Ciregan, U. Meier, and J. Schmidhuber, in *2012 IEEE conference on computer vision and pattern recognition* (IEEE, Providence, RI, 2012), p. 3642.

[5] A. Graves, A.-r. Mohamed, and G. Hinton, in *2013 IEEE international conference on acoustics, speech and signal processing* (IEEE, Vancouver, Canada, 2013), p. 6645.

[6] Y. Goldberg, Neural network methods for natural language processing, Synthesis Lectures on Human Language Technologies **10**, 1 (2017).

[7] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature **533**, 73 (2016).

[8] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[9] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[10] J. Liu, Y. Qi, Z. Y. Meng, and L. Fu, Self-learning monte carlo method, Phys. Rev. B **95**, 041101(R) (2017).

[11] L. Huang and L. Wang, Accelerated monte carlo simulations with restricted boltzmann machines, Phys. Rev. B **95**, 035105 (2017).

[12] J. Liu, H. Shen, Y. Qi, Z. Y. Meng, and L. Fu, Self-learning monte carlo method and cumulative update in fermion systems, Phys. Rev. B **95**, 241104(R) (2017).

[13] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, and T. Wongjirad, Machine learning at the energy and intensity frontiers of particle physics, Nature **560**, 41 (2018).

[14] H. Shen, J. Liu, and L. Fu, Self-learning monte carlo with deep neural networks, Phys. Rev. B **97**, 205140 (2018).

[15] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nat. Phys. **14**, 447 (2018).

[16] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[17] A. M. Palmieri, E. Kovlakov, F. Bianchi, D. Yudin, S. Straupe, J. D. Biamonte, and S. Kulik, Experimental neural network enhanced quantum tomography, Npj Quantum Inf. **6**, 1 (2020).

[18] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, A million spiking-neuron integrated circuit with a scalable communication network and interface, Science **345**, 668 (2014).

[19] S. Jutamulia and F. Yu, Overview of hybrid optical neural networks, Opt. Laser Technol. **28**, 59 (1996).

[20] Y. Abu-mostafa and D. Psaltis, Optical neural computers, Sci. Am. **256**, 88 (1987).

[21] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutiérrez, L. Pesquera, C. R. Mirasso, and I. Fischer, Photonic information processing beyond turing: An optoelectronic implementation of reservoir computing, Opt. Express **20**, 3241 (2012).

[22] F. Duport, B. Schneider, A. Smerieri, M. Haelterman, and S. Massar, All-optical reservoir computing, Opt. Express **20**, 22783 (2012).

[23] T.-Y. Cheng, D.-Y. Chou, C.-C. Liu, Y.-J. Chang, and C.-C. Chen, Optical neural networks based on optical fiber-communication system, Neurocomputing **364**, 239 (2019).

[24] S. Jiao, J. Feng, Y. Gao, T. Lei, Z. Xie, and X. Yuan, Optical machine learning with incoherent light and a single-pixel detector, Opt. Lett. **44**, 5186 (2019).

[25] X. Sui, Q. Wu, J. Liu, Q. Chen, and G. Gu, A review of optical neural networks, IEEE Access **8**, 70773 (2020).

[26] X. Guo, T. D. Barrett, Z. M. Wang, and A. I. Lvovsky, Backpropagation through nonlinear units for the all-optical training of neural networks, Photon. Res. **9**, B71 (2021).

[27] B. J. Shastri, A. N. Tait, T. F. de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, Photonics for artificial intelligence and neuromorphic computing, Nat. Photonics **15**, 102 (2021).

[28] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. Miller, and D. Psaltis, Inference in artificial intelligence with deep optics and photonics, Nature **588**, 39 (2020).

[29] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, and

A. Raja *et al.*, Parallel convolutional processing using an integrated photonic tensor core, Nature **589**, 52 (2021).

[30] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, and R. Morandotti *et al.*, 11 tops photonic convolutional accelerator for optical neural networks, Nature **589**, 44 (2021).

[31] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities, Nature **569**, 208 (2019).

[32] Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, All-optical neural network with nonlinear activation functions, Optica **6**, 1132 (2019).

[33] A. Ryou, J. Whitehead, M. Zhelyeznyakov, P. Anderson, C. Keskin, M. Bajcsy, and A. Majumdar, Free-space optical neural network based on thermal atomic nonlinearity, arXiv:2102.04464 [cs.ET] (2021).

[34] D. Woods and T. J. Naughton, Photonic neural networks, Nat. Phys. **8**, 257 (2012).

[35] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, Deep learning with coherent nanophotonic circuits, Nat. Photonics **11**, 441 (2017).

[36] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, All-optical machine learning using diffractive deep neural networks, Science **361**, 1004 (2018).

[37] W. Sung, S. Shin, and K. Hwang, Resiliency of deep neural networks under quantization, arXiv:1511.06488 (2015).

[38] K. Lu and B. E. Saleh, Theory and design of the liquid crystal tv as an optical spatial phase modulator, Opt. Eng. **29**, 240 (1990).

[39] S. E. Harris, Electromagnetically induced transparency, Phys. Today **50(7)**, 36 (1997).

[40] M. Fleischhauer, A. Imamoglu, and J. P. Marangos, Electromagnetically induced transparency: Optics in coherent media, Rev. Mod. Phys. **77**, 633 (2005).

[41] H. J. Metcalf and P. van der Straten, Laser cooling and trapping of atoms, JOSA B **20**, 887 (2003).

[42] S. Zhang, J. F. Chen, C. Liu, S. Zhou, M. M. T. Loy, G. K. L. Wong, and S. Du, A dark-line two-dimensional magneto-optical trap of 85rb atoms with high optical depth, Rev. Sci. Instrum **83**, 073102 (2012).

[43] N. Matsumoto, T. Inoue, T. Ando, Y. Takiguchi, Y. Ohtake, and H. Toyoda, High-quality generation of a multispot pattern using a spatial light modulator with adaptive feedback, Opt. Lett. **37**, 3135 (2012).

[44] D. Kim, A. Keesling, A. Omran, H. Levine, H. Bernien, M. Greiner, M. D. Lukin, and D. R. Englund, Large-scale uniform optical focus array generation with a phase spatial light modulator, Opt. Lett. **44**, 3178 (2019).

[45] F. Nogrette, H. Labuhn, S. Ravets, D. Barredo, L. Béguin, A. Vernier, T. Lahaye, and A. Browaeys, Single-Atom Trapping in Holographic 2d Arrays of Microtraps with Arbitrary Geometries, Phys. Rev. X **4**, 021034 (2014).

[46] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.15.054034 for extra material on the detailed fitting results, input weight generation, measured nonlinear activation functions, neural-network training details, weighted Gerchberg-Saxton algorithm, and calculation of the capacity of the spatial light modulator.

[47] The MNIST database of handwritten digits, http://yann.lecun.com/exdb/mnist/,.

[48] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, arXiv:1708.07747 (2017).

[49] R. Di Leonardo, F. Ianni, and G. Ruocco, Computer generation of optimal holograms for optical trap arrays, Opt. Express **15**, 1913 (2007).