Fourier-space Diffractive Deep Neural Network

Tao Yan,[1,*] Jiamin Wu,[1,*] Tiankuang Zhou,[1,2,*] Hao Xie,[1] Feng Xu,[3]
Jingtao Fan,[1] Lu Fang,[2] Xing Lin,[1,4,†] and Qionghai Dai[1,‡]

[1]*Department of Automation, Tsinghua University, Beijing, 100084, People's Republic of China*
[2]*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, People's Republic of China*
[3]*School of Software, Tsinghua University, Beijing 100084, People's Republic of China*
[4]*Department of Electrical and Computer Engineering, University of California, Los Angeles, California 90095, USA*

(Received 2 April 2019; revised manuscript received 6 June 2019; published 9 July 2019)

In this Letter we propose the Fourier-space diffractive deep neural network (F-D^{2}NN) for all-optical image processing that performs advanced computer vision tasks at the speed of light. The F-D^{2}NN is achieved by placing the extremely compact diffractive modulation layers at the Fourier plane or both Fourier and imaging planes of an optical system, where the optical nonlinearity is introduced from ferroelectric thin films. We demonstrated that F-D^{2}NN can be trained with deep learning algorithms for all-optical saliency detection and high-accuracy object classification.

DOI: 10.1103/PhysRevLett.123.023901

Performing modern computer vision tasks, especially with the resurgent deep learning algorithms, requires large-scale image processing and an increase in demand for computational resources [1–3]. However, Moore's law for electronic computing is continuously slowing down, and the scale of an electronic transistor is approaching the physical limit [4,5]. Optical computing offers low power consumption, light-speed processing, and high-throughput capability, which possesses the inherent potential to serve as significant support for high-performance computing [6–8]. Recent research on metamaterials and nanophotonics have paved the way to the all-optical signal processing and integrated optical circuit, but most of them can only implement simple image processing operations, such as differentiation, integration, and convolution [9–17]. Building the optical neural network has been proven effective for solving more complex computational problems, including image classification and speech recognition [7,18–25]. In this line of work, an all-optical machine learning framework, termed as a diffractive deep neural network (D^{2}NN) [18], was recently introduced to perform the light-speed classification and high-resolution imaging. However, it only operates in real space and has limited capability to address more advanced computer vision tasks. This Letter proposes the Fourier-space D^{2}NN (F-D^{2}NN) and demonstrates the success of its application for salient object detection. Besides, we conduct various numerical evaluations and show that the classification accuracy and robustness of D^{2}NN is significantly improved by training it in Fourier space and including the optical nonlinearity.

Salient object detection refers to the detection and segmentation of salient objects in natural scenes, which has attracted great interest in the computer vision community due to its capability of finding objects or regions for efficiently representing a scene [26–30]. Different from the saliency prediction [31,32] that tries to predict human eye fixation over an image at first glance, we focus on highlighting the most important object regions in an image. As the number of sensor measurements for capturing larger field of view and higher resolution images in both photography and microscopy areas [33,34] has continuously increased, using only electronic computing has placed a major bottleneck for processing such high-throughput data efficiently. This Letter introduces the all-optical salient object detection by simply placing the D^{2}NN at the Fourier plane of an optical system, where the sensor directly measures the salient object detection results. The proposed architecture is set to work under a visible wavelength of 532 nm, which offers the network with a high resolution and integration. The quantitative evaluation of the all-optical saliency detection for both macroscale scenes and microscopic samples demonstrates the effectiveness of the proposed approach.

The nonlinear activation function is one of the crucial components in artificial neural networks. We propose to employ a specific type of ferroelectric thin film, i.e., photorefractive crystal (SBN:60) [35–37] as an optical nonlinearity layer and present different system configurations to incorporate the nonlinearity layers physically. The photorefractive crystal's (SBN:60) nonlinear material modifies its refractive index with the changing of light intensity in the medium, which performs the complex activation function for the output complex field of a neuron. Such complex threshold preserves the light efficiency in contrast to the intensity threshold. Training the diffractive modulation layers of D^{2}NN in both Fourier and real spaces with nonlinear activation layers enables the fitting of more complex mapping functions, which significantly improves

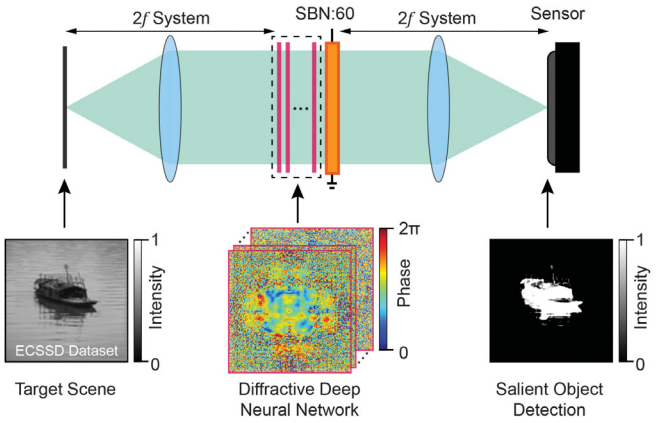

FIG. 1. Salient object detection with Fourier-space diffractive deep neural network (F-D^{2}NN). The F-D^{2}NN optical image processing module is formed by inserting the D^{2}NN along with a photorefractive crystal (SBN:60) at the Fourier plane of an optical system under coherent light. F-D^{2}NN can achieve all-optical segmentation of the salient objects for the target scene after deep learning design of modulation layers.

its performance on the task of classification. With the presented optical nonlinearity, we also demonstrate that Fourier-space D^{2}NN can achieve higher classification accuracy and much more compact structure than real-space D^{2}NN [18] under the visible spectrum.

The framework of the proposed Fourier-space D^{2}NN is demonstrated in Fig. 1. Under the coherent illumination, the complex optical field of a target scene $\mathbf{U}_0$ is Fourier transformed and fed into the input diffractive layer of D^{2}NN through a $2f$ optical system: $\hat{\mathbf{U}}_0 = \mathbf{F}\mathbf{U}_0$, where $\mathbf{F}$ is the Fourier transform matrix. The D^{2}NN performs a complex transform function $\hat{\mathbf{M}}$ in the Fourier space for the input optical field after training with deep learning algorithms, which computes the output optical field $\hat{\mathbf{U}}_1$ of a network through optical diffractions layer by layer: $\hat{\mathbf{U}}_1 = \hat{\mathbf{M}}\hat{\mathbf{U}}_0 = \hat{\mathbf{M}}\mathbf{F}\mathbf{U}_0$. In this Letter, we target to train phase-only diffractive modulation layers. By attaching a photorefractive crystal, the optical field after propagating through nonlinear material with complex activation function φ can be formulated as $\hat{\mathbf{U}}_2 = \varphi(\hat{\mathbf{U}}_1) = \varphi(\hat{\mathbf{M}}\mathbf{F}\mathbf{U}_0)$, which is then Fourier transformed back to the real space with another $2f$ optical system and the sensor measures the intensity distribution of the optical field at the output plane: $\mathbf{O} = |\mathbf{F}\hat{\mathbf{U}}_2|^2 = |\mathbf{F}\varphi(\hat{\mathbf{M}}\mathbf{F}\mathbf{U}_0)|^2$.

During the training, input images are encoded into the amplitude of complex field $\mathbf{U}_0$ and sensor measurements $\mathbf{O}$ are computed with the formulated forward model, which are then used to calculate errors with respect to the ground truth targets $\mathbf{O}_{gt}$. The loss function with mean squared error evaluation criterion can be defined as $e(\hat{\mathbf{M}}) = \|\mathbf{O} - \Gamma(\mathbf{O}_{gt})\|$, where Γ represents the operator of reversing coordinate axes due to the using of two optical Fourier transforms. The resulting errors are backpropagated to iteratively update the phase modulation coefficients of the diffractive neural network and eventually minimize the loss function. After the training, the F-D^{2}NN framework is fixed and diffractive modulation coefficients are determined, which can be used for physical fabrication. More details of network training are provided in the Supplemental Material [38].

Compared with the real-space D^{2}NN [18], our framework is more natural to preserve the spatial correspondence by incorporating a dual $2f$ optical system, which facilitates those tasks that require an image-to-image mapping. The transform function is learned in Fourier space since every different object will have a very different Fourier transform pattern. To demonstrate, we take the first step to address the task of salient object detection all optically with the proposed F-D^{2}NN framework. In the proposed framework, each input point of the target scene is fully connected to all of the neurons in the D^{2}NN and maps to the corresponding location on the detector, where the D^{2}NN statistically learns to perform saliency filtering on different spatial frequency components of the target scene after the training, similar to the spectral residual approach proposed in Ref. [29]. Besides, the employed nonlinear layer can facilitate the nonlinear representation of the target scene to better capture its structures than a linear representation. These underline physical principles place constraints on the searching space of network and guarantee the reliable estimation of visual saliency.

We validated the proposed F-D^{2}NN for performing an all-optical salient object detection by demonstrating its application in cell segmentation. For this task, the phase-only modulation layers were designed by training a five-layer F-D^{2}NN with the pathology slide images from National Cancer Institute GDC Data Portal [39]. During the implementation, the numerical aperture of the dual $2f$ system was set to match the neuron size of the D^{2}NN so that the optical diffraction used for computing can be captured for generating saliency maps on the sensor. We first demonstrated the unit-magnification system with the network setting details in the Supplemental Material [38]. To facilitate the practical fabrication of phase modulation layers with direct femtosecond laser writing [40,41], the distance between successive layers was optimized and set to be 100 μm that offers the network a highly compact structure. The network was trained using pathology slide images of cell type 1 with 2750 images and 250 validation images. The ground truth saliency detection results used for training and evaluation are obtained with state-of-the-art algorithms, including cosaliency detection [27] and robust background detection (RBD) algorithms [30].

We blindly tested the trained F-D^{2}NN with cell type 1 from testing dataset (500 images) as well as the cell type 2 (250 images) and cell type 3 (250 images) for evaluating the generalization of the network, as demonstrated in Fig. 2. The ground truth results in this implementation were obtained with the cosaliency detection algorithm. The

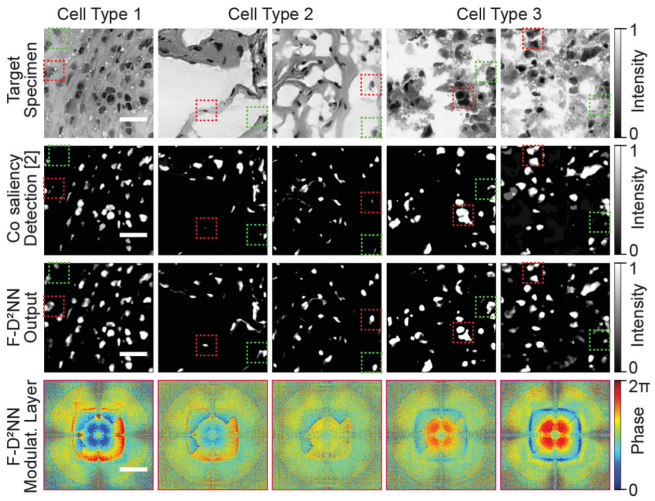

FIG. 2. Cell segmentation with F-D^{2}NN on testing images. The F-D^{2}NN is trained (Cell Type 1) and tested (Cell Type 1, 2 and 3) with pathology slide images from National Cancer Institute GDC Data Portal [39] for salient object detection and applying it to cell segmentation. After the training, the patterns of F-D^{2}NN modulation layers (fourth row) are fixed and used to perform all-optical salient object detection (third row, intensity represents the saliency level), which can achieve comparable accuracy compared with the state-of-the-art cosaliency detection algorithm [27] (second row) for the input images (first row). (Scale bar: 200 μm for the first to the third row, 400 μm for the fourth row. Cell scales are optically magnified).

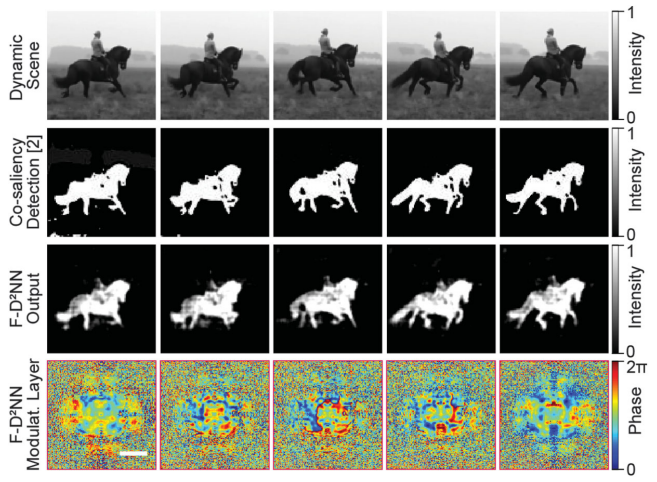

FIG. 3. Video salient object detection with F-D^{2}NN. F-D^{2}NN is trained with CIFAR-10 database and tested with online video sequence [42]. F-D^{2}NN modulation layers are converged to the patterns as shown in the fourth row, which are then used for performing all-optical salient detection (third row) for input image sequences (first row). (Scale bar: 80 μm).

results demonstrated that F-D^{2}NN successfully learned to perform visual saliency detection (third row) for the target specimen (first row) and obtained comparable results with respect to the ground truths (second row). As expected, the saliency detection results segmented out the cells and revealed their shapes and locations. The cosaliency detection approach proposed in Ref. [27] imposes a location constraint that the salient object has a higher probability of being in the center of a human-made image. However, the trained F-D^{2}NN is able to learn and average various spatial features at different locations of the training images; thus the detection and segmentation results are even better than the cosaliency detection approach on the peripheral regions, as marked out in Fig. 2. For the quantitative evaluation, we calculate the precision-recall (PR) curve [26] by comparing the saliency detection results of testing images with respect to the ground truths. The maximum F measures of average PR curve over three testing datasets for cell type 1, 2, and 3 are 0.613, 0.653, and 0.451, respectively. The proposed F-D^{2}NN also works for the magnification or demagnification system, as demonstrated in Fig. S1. The same conclusion can be obtained by using $4f$ system instead of dual $2f$ system or changing the ground truth reconstruction algorithm from cosaliency detection to RBD algorithm as shown in the Figs. S2 and S3, respectively.

We further demonstrate the generality of F-D^{2}NN for video saliency detection of dynamic natural scenes at the macroscopic scale. With the network setting details in the Supplemental Material [38] and using the cosaliency detection results as the ground truth, Fig. 3 shows the testing of online video sequence [42] after training five layers F-D^{2}NN with CIFAR-10 dataset [43]. The trained diffractive modulation layers are demonstrated in the last row of Fig. 3, with which the input video sequence (first row) was tested frame by frame (240 frames in total) to generate the saliency video sequence (third row). The Supplemental Video 1 shows the result of whole video sequences, and the maximum F measure on the averaged PR curve with respect to the cosaliency detection algorithm (second row) is calculated to be 0.726. The results demonstrate the potential application of the proposed approach for high-speed and robustness saliency detection of dynamic scenes. To quantitatively evaluate the performance of the proposed all-optical computing framework with respect to the state-of-the-art visual saliency detection algorithms, we also test the trained model with benchmark datasets including the DAVIS [44] and ECSSD [45]. The corresponding results are shown in the Figs. S4 and S5 as well as Supplemental Video 2, where the maximum F measures on the results of the proposed approach are 0.498 and 0.423, respectively.

Furthermore, we compared the Fourier- and real-space D^{2}NN under different configurations for both tasks of saliency detection and object classification, as shown in Fig. 4. The phase modulation range of each neuron was constrained to a certain range as details in Fig. S6. We trained both five-layer Fourier and real-space D^{2}NN on CIFAR-10 dataset "Cat" category and tested the trained model on "Horse" category. As expected, the real-space D^{2}NNs with

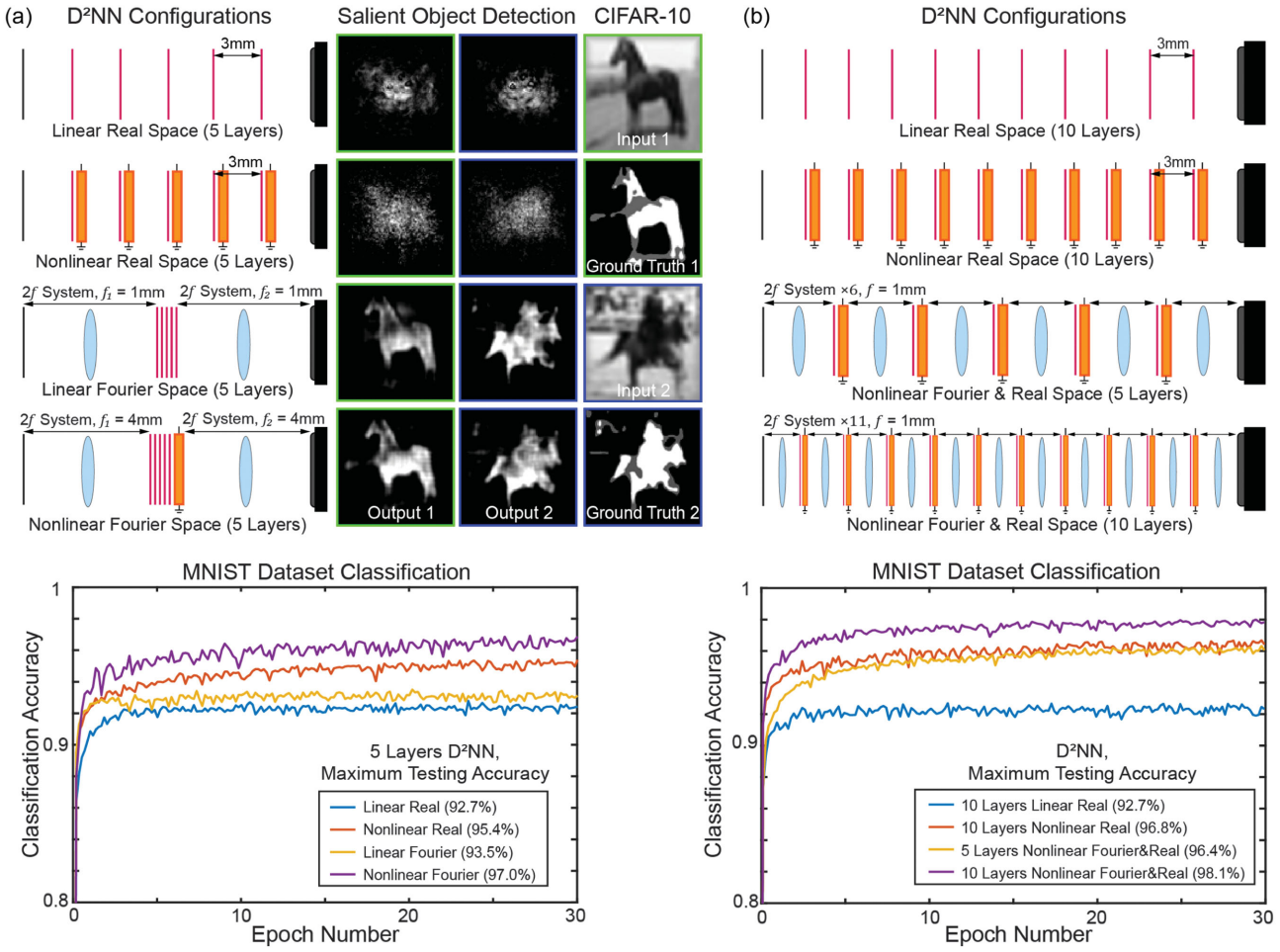

FIG. 4. The comparisons between Fourier- and real-space D^{2}NN for salient object detection and object classification. (a) Real-space D^{2}NN cannot achieve correct localization and segmentation of the target while Fourier-space D^{2}NN successfully detects the salient object and performs the segmentation. The transforming of D^{2}NN from real-space into Fourier-space and the incorporating of the nonlinear optical layer significantly reduce its thickness and improve the classification accuracy. (b) The classification accuracy of linear real-space D^{2}NN saturates with the increasing of layer number. However, with the incorporating of nonlinear optical layers, the classification accuracy increases with the increasing of layer number, and the hybrid method of Fourier and real-space D^{2}NN can further improve the accuracy.

or without the nonlinearity fail on the task of salient object detection due to its difficulty of finding spatial correspondences, while F-D^{2}NN successfully performs the detection by training and testing on the same datasets that real-space D^{2}NN used [Fig. 4(a)]. The maximum F measure of saliency detection on the testing dataset for the nonlinear F-D^{2}NN is 0.641, which outperformed the linear configuration (0.634). For the task of object classification, we use the MNIST handwritten digit dataset like Ref. [18] for quantitatively evaluating different D^{2}NN configurations and calculating the classification accuracy. The photorefractive crystal is incorporated as the nonlinearity layer by setting the thickness to be 1 mm for generating enough strength of phase variation (details in the Supplemental Material [38]). We found that training linear D^{2}NN in Fourier space not only reduces the thickness of a five-layer D^{2}NN (layer distance from 3 mm to 100 μm) but also improves its classification accuracy (from 92.7 to 93.5%). By using a single nonlinearity layer behind modulation layers to the Fourier-space configuration, the classification accuracy was significantly improved to 97.0% with the optimal layer distance of 100 μm. The resulting compact network structure facilitates network fabrication and physical experiments. Increasing the number of nonlinearity layers achieves comparable classification accuracy (Fig. S7), but it decreases the performance of saliency detection (Fig. S2) and reduces the feasibility for physical implementation. Therefore, the nonlinear F-D^{2}NN in this Letter is configured with a single nonlinearity layer. With a comparable system length of nonlinear F-D^{2}NN, the five-layer real-space D^{2}NN configured with five nonlinearity layers has the classification accuracy of 95.4%. The classification accuracy of linear real-space D^{2}NN saturates with the increasing of layer number from five to ten, but with the nonlinearity layer, ten layers of nonlinear D^{2}NN at real space

has the classification accuracy of 96.8%. By placing the diffractive modulation layers on both Fourier space and real space, a five-layer nonlinear hybrid D^2NN configuration has the classification accuracy of 96.4%, which can be improved to 98.1% with ten layers. All convergence plots of the described D^2NN configurations on the MNIST blind testing dataset are demonstrated in Fig. 4 (bottom row). In the Supplemental Material [38], we further evaluated the robustness of F-D^2NN with respect to the fabrication and alignment errors on performing the tasks of classification (Fig. S8) and saliency detection (Fig. S9) and demonstrated the feasibility of the proposed architecture for physical experiments.

To conclude, we have shown that the F-D^2NN can be applied to achieve advanced image processing and computer vision tasks at the speed of light. We validated the effectiveness of the proposed approach for high-accuracy visual saliency detection and object classification through various numerical experiments. Similar to coded aperture imaging techniques, such as phase contrast microscopy, our method can be implemented as an intelligent optical filter and adapted to different imaging systems including commercial microscopes and cameras.

The executable codes and datasets in this Letter are available upon reasonable request.

This Letter was supported by the Project of Beijing Municipal Commission of Science and Technology (No. Z181100003118014) and National Natural Science Foundation of China (No. 61327902 and No. 61722209).

*T. Y., J. W. and T. Z. contributed equally to this Letter.
†linx2019@tsinghua.edu.cn
‡qhdai@tsinghua.edu.cn

[1] Y. LeCun, Y. Bengio, and G. Hinton, Nature (London) **521**, 436 (2015).
[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., New York, USA, 2012), pp. 1097–1105.
[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, Nature (London) **529**, 484 (2016).
[4] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, *Unconventional Computing: A Volume in the Encyclopedia of Complexity and Systems Science*, 2nd ed. (Springer, New York, 2018), Vol. 83.
[5] J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, Phys. Rev. Applied **7**, 034013 (2017).
[6] D. R. Solli and B. Jalali, Nat. Photonics **9**, 704 (2015).
[7] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund *et al.*, Nat. Photonics **11**, 441 (2017).
[8] D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, Nature (London) **343**, 325 (1990).
[9] A. Silva, F. Monticone, G. Castaldi, V. Galdi, A. Alù, and N. Engheta, Science **343**, 160 (2014).
[10] N. M. Estakhri, B. Edwards, and N. Engheta, Science **363**, 1333 (2019).
[11] H. Kwon, D. Sounas, A. Cordaro, A. Polman, and A. Alù, Phys. Rev. Lett. **121**, 173004 (2018).
[12] T. Zhu, Y. Zhou, Y. Lou, H. Ye, M. Qiu, Z. Ruan, and S. Fan, Nat. Commun. **8**, 15391 (2017).
[13] M. Ferrera, Y. Park, L. Razzari, B. E. Little, S. T. Chu, R. Morandotti, D. Moss, and J. Azaña, Nat. Commun. **1**, 29 (2010).
[14] C. Guo, M. Xiao, M. Minkov, Y. Shi, and S. Fan, Optica **5**, 251 (2018).
[15] A. Pors, M. G. Nielsen, and S. I. Bozhevolnyi, Nano Lett. **15**, 791 (2015).
[16] J. Xu, X. Zhang, J. Dong, D. Liu, and D. Huang, Opt. Lett. **32**, 1872 (2007).
[17] A. Chizari, S. Abdollahramezani, M. V. Jamali, and J. A. Salehi, Opt. Lett. **41**, 3451 (2016).
[18] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, Science **361**, 1004 (2018).
[19] J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, Optica **5**, 756 (2018).
[20] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, Optica **5**, 864 (2018).
[21] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, Sci. Rep. **8**, 12324 (2018).
[22] I. Chakraborty, G. Saha, A. Sengupta, and K. Roy, Sci. Rep. **8**, 12980 (2018).
[23] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Sci. Rep. **7**, 7430 (2017).
[24] M. Hermans, M. Burm, T. Van Vaerenbergh, J. Dambre, and P. Bienstman, Nat. Commun. **6**, 6729 (2015).
[25] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, Nat. Commun. **4**, 1364 (2013).
[26] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, IEEE Trans. Image Process. **24**, 5706 (2015).
[27] H. Fu, X. Cao, and Z. Tu, IEEE Trans. Image Process. **22**, 3766 (2013).
[28] W. Wang, J. Shen, and L. Shao, IEEE Trans. Image Process. **27**, 38 (2018).
[29] X. Hou and L. Zhang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, 2007), pp. 1–8.
[30] W. Zhu, S. Liang, Y. Wei, and J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, 2014), pp. 2814–2821.
[31] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, 2016), pp. 598–606.
[32] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)* (IEEE, Piscataway, 2016), pp. 3488–3493.
[33] D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. Vera, and S. D. Feller, Nature (London) **486**, 386 (2012).
[34] G. Zheng, R. Horstmeyer, and C. Yang, Nat. Photonics **7**, 739 (2013).

[35] L. Waller, G. Situ, and J. W. Fleischer, Nat. Photonics **6**, 474 (2012).
[36] D. N. Christodoulides, T. H. Coskun, M. Mitchell, and M. Segev, Phys. Rev. Lett. **78**, 646 (1997).
[37] R. W. Boyd, *Nonlinear Optics* (Elsevier, Amsterdam, 2003).
[38] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.123.023901 for details about the network training, optical nonlinearity property, as well as the ablation and sensitivity analysis of the network with different configurations and evaluation datasets.
[39] National Cancer Institute GDC Data Portal, https://portal.gdc.cancer.gov.
[40] Y.-L. Zhang, Q.-D. Chen, H. Xia, and H.-B. Sun, Nano Today **5**, 435 (2010).
[41] A. Ródenas, M. Gu, G. Corrielli, P. Paiè, S. John, A. K. Kar, and R. Osellame, Nat. Photonics **13**, 105 (2019).
[42] Online Horse Video, http://3g.163.com/v/video/VXI1LO67N.html.
[43] A. Krizhevsky and G. Hinton, Learning multiple layers of features from tiny images, University of Toronto Technical Report, Citeseer, 2009, http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.
[44] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, 2016), pp. 724–732.
[45] J. Shi, Q. Yan, L. Xu, and J. Jia, IEEE Trans. Pattern Anal. Mach. Intell. **38**, 717 (2016).