Advances in Optics and Photonics

# At the intersection of optics and deep learning: statistical inference, computing, and inverse design

Deniz Mengu,[1,2,3] Md Sadman Sakib Rahman,[1,2,3] Yi Luo,[1,2,3] Jingxi Li,[1,2,3] Onur Kulce,[1,2,3] and Aydogan Ozcan[1,2,3,*]

[1]Electrical and Computer Engineering Department, University of California, Los Angeles, CA, 90095, USA
[2]Bioengineering Department, University of California, Los Angeles, CA, 90095, USA
[3]California NanoSystems Institute (CNSI), University of California, Los Angeles, CA, 90095, USA
*Corresponding email: ozcan@ucla.edu

Received December 2, 2021; revised April 3, 2022; accepted April 4, 2022; published 19 May 2022

Deep learning has been revolutionizing information processing in many fields of science and engineering owing to the massively growing amounts of data and the advances in deep neural network architectures. As these neural networks are expanding their capabilities toward achieving state-of-the-art solutions for demanding statistical inference tasks in various applications, there appears to be a global need for low-power, scalable, and fast computing hardware beyond what existing electronic systems can offer. Optical computing might potentially address some of these needs with its inherent parallelism, power efficiency, and high speed. Recent advances in optical materials, fabrication, and optimization techniques have significantly enriched the design capabilities in optics and photonics, leading to various successful demonstrations of guided-wave and free-space computing hardware for accelerating machine learning tasks using light. In addition to statistical inference and computing, deep learning has also fundamentally affected the field of inverse optical/photonic design. The approximation power of deep neural networks has been utilized to develop optics/photonics systems with unique capabilities, all the way from nanoantenna design to end-to-end optimization of computational imaging and sensing systems. In this review, we attempt to provide a broad overview of the current state of this emerging symbiotic relationship between deep learning and optics/photonics. 

https://doi.org/10.1364/AOP.450345

# At the intersection of optics and deep learning: statistical inference, computing, and inverse design

Deniz Mengu, Md Sadman Sakib Rahman, Yi Luo, Jingxi Li, Onur Kulce, and Aydogan Ozcan

## 1. INTRODUCTION

Historically, optical devices and systems have been extensively used to sense/detect, communicate, store, and display information. Although the idea of leveraging the speed of photons and the inherent parallelism of optics for general-purpose, low-latency, energy-efficient computation has been highly appealing for decades, it has yet to find widespread acceptance, affecting our everyday lives, which continue to be dominated by electronic computers [1]. In contrast to optical computing, on the other hand, deep learning and artificial neural networks (ANNs) have already established their widespread recognition as the mainstream algorithmic tool for learning, information processing, and statistical inference. With the massively growing amounts of data and the complexity of the accompanying artificial intelligence (AI) algorithms in a range of demanding applications, e.g., autonomous driving, robotics, remote sensing, defense, Internet-of-Things (IoT), the emerging needs of modern-day computing hardware on speed, energy-efficiency, and form factor have started to point beyond the reach of electronic computers; this, in turn, presents another opportunity for optics and photonics to contribute new solutions to our global computing needs.

Traditional electronic computers are mainly built upon the Von Neumann architecture [2], where the memory and the processor are two separate units communicating over a bus at a limited data rate, operating in a sequential manner. Although this architecture had not posed major problems in its early stages, when the processing unit was the slowest link in the chain, starting with the mid-1990s, the clock speed of processors surpassed the speed of memory, i.e., the execution became faster than the data feed, creating the problem known as the "von Neumann bottleneck" [2]. In addition, the exponential gain in the computational capacity of electronics predicted by Moore's law has dramatically slowed down recently, as the size of transistors approached close to their physical limits. One remedy for these architectural inefficiencies and decelerated developments in transistor fabrication has been found in the use of graphics processing units (GPUs) for general-purpose computing. Together with the availability of massive data repositories, the immense processing power and parallelism of modern GPUs have been one of the major driving forces behind the rise and success of deep learning and ANNs. Currently, GPUs represent one of the most mature and accessible computing hardware for the training and execution of ANNs, which require massive amounts of data to be rapidly processed. On the other hand, the enormous processing capacity of high-end GPUs is accompanied by their high power needs, bulky form factor, and relatively high processing latency. These factors make it challenging to deploy machine learning algorithms based on complex ANNs on low-power, size- and/or cost-limited systems, such as mobile devices and edge computing applications, e.g., cameras, autonomous vehicles, and IoT peripherals [3].

Despite the maturity of electronic computing technologies including GPUs, optical networks and photonic circuits might play a major role in the future of mobile AI,

edge computing, and other machine-learning-related applications; optics and photonics present the potential to offer massively parallel, fast operation with scalable, small form-factor devices that have very low power consumption. For example, the feedforward ANN models, already trained digitally using GPUs, can be deployed as all-optical and/or hybrid (optical–electronic) machine learning platforms to carry out feedforward inference tasks with low latency and low power consumption, using small-form-factor computing devices. In particular, inference tasks in visual computing applications, where the information is already in the optical domain (e.g., a scene) are well suited for exploiting optical computing techniques, because the fundamental building blocks of ANNs, such as convolutions, matrix–vector multiplications, and various transformations, can be executed all-optically to be completed at the speed of light propagation, as a byproduct of diffraction and light–matter interactions.

Although a tremendous amount of effort was devoted to developing optical neural networks and related computing schemes, especially during the 1980s, these earlier studies did not result in practical applications due to inadequate data, shortcomings of available fabrication techniques, and limited access to state-of-the-art computing technologies in these earlier decades. The wide availability of GPUs and significant advances in nano-fabrication techniques and optical materials [4,5] during the past few decades have refueled the research on optical neural network and photonic computing techniques. Once some of these design ideas eventually mature into practical technologies, they could potentially drive a paradigm shift in implementations of machine learning models and expand the reach of modern AI. Similar to the possible advantages offered by optical computing systems for AI applications, recent developments in various branches of optics and photonics have also greatly benefited from the advances in deep neural networks and other machine learning tools. For instance, deep learning and neural networks have already been shown to produce state-of-the-art results for computational inverse problems in many optical imaging and sensing applications, e.g., microscopy [6–10], holography [11–15], quantitative phase imaging (QPI) [14,16], among many others [17–33].

In addition to these applications, the data-driven nature of deep learning and its inference capabilities are exploited to solve challenging, task-specific inverse optical design problems for various applications, including metamaterials, nanophotonics, free-form optics, and imaging. Traditional solutions of hardware design problems in optics and photonics have, almost exclusively, been driven by the physical laws, e.g., Maxwell's equations, describing the forward model of the underlying wave phenomena. However, relying heavily on analytical closed-form solutions often restricts the hardware design space to mathematically tractable representations, which necessitates resource-intensive numerical methods. Although the inverse hardware designs in various fields, e.g., nanophotonics and metamaterials, have greatly benefited from the adaptation of numerous optimization schemes from greedy search algorithms to metaheuristic approaches such as evolutionary algorithms, these numerical optimization procedures often require solving the forward model recursively with many repetitions, making them relatively slow. The universal function approximation power of deep neural networks [34], on the other hand, presents new avenues for inverse optical design problems. In the field of nanophotonics [35], for instance, deep neural networks can be used as surrogate models, also known as metamodels. In this approach, a deep neural network is trained to solve for the forward physical transformation, i.e., given a set of design parameters as inputs, the trained deep neural network approximates the system response at the output in a much faster manner compared with rigorous solvers of Maxwell's equations [36,37]. The exact opposite functionality is also offered by deep-learning-based solutions, meaning that deep networks can be trained to directly

infer an optical hardware design to yield a desired electromagnetic (EM) response at the output [38].

In other fields of optics, where modeling of light as a scalar wave field is adequate, deep learning has also been utilized to tailor the light–matter interactions over a series of diffractive surfaces in a task-specific and data-driven manner [39,40]. These multilayer diffractive platforms have been shown to provide non-intuitive, task-specific solutions to inverse optical design problems, e.g., spatially controlled wavelength demultiplexing [41] and lensless all-optical pulse shaping [42]. As another example, in imaging and microscopy-related applications, deep neural networks and optical system architecture in front of the optoelectronic sensors have been co-designed in a unified computational framework to optimize the imaging system performance in various metrics such as spatial resolution, dynamic range, and depth of field (DOF) [43,44]. It has also been shown that the deep-learning-based software–hardware co-design can be used to mitigate aberrations and other forms of imperfections due to, e.g., fabrication errors in computational imaging and sensing systems [45,46].

As highlighted through these examples from the recent literature, there is an emerging symbiotic relationship between optics/photonics and deep learning that is immensely beneficial for both research fields. Scientific and engineering advances at this intersection of optics/photonics and deep learning are emerging at an unprecedented speed and can not only leapfrog our design methods in optics, but can also provide highly parallelized, scalable, and low-power, extremely fast computing platforms to further expand the capabilities and application areas of deep neural networks. To highlight these exciting opportunities and emerging advances, in this article we provide a timely review of the use of deep learning in optics and photonics for statistical inference, computing, and inverse design applications.

## 2. BACKGROUND ON OPTICAL COMPUTING AND NEURAL NETWORKS

We start this section with a brief historic overview of deep learning and neural networks.

### 2.1. Deep Learning and Neural Networks

Although the term “deep learning” has recently become widely popular, the fundamental ideas represented by this phrase date back to the 1940s, e.g., “all-or-none” McCulloch–Pitts neuron [47]. At that time, the error-backpropagation algorithm had not been discovered yet; hence, the weights had to be manually adjusted by a human and the model was entirely linear, trying to categorize the input into two classes, as positive or negative. Rosenblatt pioneered the first perceptron model that can learn the weights to categorize given inputs [48]. A modified version of the stochastic gradient descent, one of the most widely used optimization schemes in modern deep learning, was first used to train an adaptive switching circuit, called ADALINE (Adaptive Linear).

Although the earliest research efforts in the field were intended to computationally mimic the learning functionality of the brain, the proposed models had very little correspondence with the biological structure and functions inside the human brain; the name “artificial neural network” can therefore be considered a metaphor. Even today, very little is known about the inner principles, computational paths, and the nervous system inside the brain. Therefore, neuroscience and the actual biological infrastructure of the brain have not been (and perhaps will never be) a strict guide for the evolution of modern deep learning techniques. The most prevailing analogy between the nervous system and ANNs is that in both, a large number of simple computational units can

show intelligence when they interact and work collaboratively with each other. Despite such rough guidelines, deep learning, today, is a field of research primarily focused on developing computational systems that can learn to solve demanding tasks requiring intelligence.

#### 2.1a. Universal Approximation Theorem

Deep learning has been a field driven mostly by empirical evidence and numerical experimentation. The theoretical proofs on the universal approximation capabilities of feedforward networks with hidden layers [49,50] were published in 1989, almost 3 years after Rumelhart *et al.* had managed to train a neural network with error-backpropagation [51]. Nevertheless, the universal approximation theorem can be considered as one of the important mathematical foundations behind the success of the modern deep learning frameworks from a theoretical perspective.

As constructed by Refs. [49,50], the theorem states that a feedforward neural network with at least one hidden layer and a linear output layer can approximate any continuous function from a finite-dimensional, closed, and bounded space to another one with any desired error tolerance, given that the network has enough hidden units each with a "squashing" nonlinear activation function. As defined in Ref. [50], a "squashing" nonlinear function $\psi(x)$ maps the set of real numbers, $\mathbb{R}$, onto the interval [1] such that if it is non-decreasing, $\lim_{x\to\infty}\psi(x) = 1$ and $\lim_{x\to-\infty}\psi(x) = 0$. Some examples of squashing functions are: (1) the indicator/threshold or unit step function, $\psi(x) = 0_{\{x<0\}} + 1_{\{0\le x\}}$, (2) ramp function, $\psi(x) = 0_{\{x\le 0\}} + x_{\{0\le x\le 1\}} + 1_{\{1\le x\}}$, and (3) the cosine squasher, $\psi(x) = 0_{\{x<-\pi/2\}} + \frac{1}{2}\left(1 + \cos\left(x + \frac{3\pi}{2}\right)\right)_{\{-\pi/2\le x\le\pi/2\}} + 1_{\{x>\pi/2\}}$. The well-known sigmoid function, $\psi(x) = \frac{1}{1+e^{-x}}$, is also a member of this family of functions [49].

Although these earlier works covered only a limited set of nonlinear activation functions that are bounded, the universal approximation theorem has been further extended to a wider set of activation functions [52], including the unbounded rectified linear unit (ReLU) function that is used extensively in modern deep neural networks. With deep neural networks already granted a wide acceptance as the state-of-the-art information processing architecture, the research on universal approximation theorem is still active, mostly focusing on exploring and expanding the class of nonlinear activation functions that can provide theoretical guarantees on the approximation capabilities and error bounds of deep neural networks [53]. Research along the lines of Refs. [52,53] can also be crucial for the future of optical neural networks and photonic processors due to the restricted space of nonlinearities that can be created through optical materials and devices. For instance, Ref. [53] introduces truncated power functions such as $f(x) = 0_{\{x\le 0\}} + x^2_{\{x>0\}}$ as a nonlinear functional form that can serve as an activation function in a universal approximator. This function, $f(x)$, might be of particular importance for optical systems, as photodetectors generate optical signals proportional to the incident light intensity, which itself is proportional to the field-amplitude squared. Another recent research that is closely related to coherent optical neural networks [39] has been published by Voigtlaender [54] investigating the universal approximation capabilities of complex-valued neural networks. The author found that, unlike real-valued neural network architectures, the desired features of nonlinear activation functions for providing universal approximation guarantees in complex-valued networks depend on the depth of the neural network. For example, in the case of complex-valued neural networks with at least two hidden layers, a broad class of nonlinear activation functions, except the polynomials and holomorphic/antiholomorphic functions, can theoretically provide universal approximation property [54].

*2.1b. Feedforward Neural Networks*

The goal of any feedforward neural network model is to approximate a desired function $g$. If, for example, the machine learning task at hand is to design a classifier, then $\boldsymbol{y} = g(\boldsymbol{x})$ is a desired mapping that assigns a category $y$ to an input $\boldsymbol{x}$. Deep learning aims to find an approximation $f(\boldsymbol{x})$ that satisfies $f(\boldsymbol{x}) \approx g(\boldsymbol{x})$ for any given input $\boldsymbol{x}$. In the case feedforward networks, $f$ is defined over an acyclic computational graph with each layer, $i$, corresponding to a function $f^{(i)}$ such that $f = f^{(L)}(f^{(L-1)}(\ldots f^{(1)}(\boldsymbol{x})))$, where $L$ is the number of successive layers in the graph. According to the universal approximation theorem, a neural network with at least one hidden layer, i.e., $f = \boldsymbol{W}f^{(1)}(\boldsymbol{x}) + \boldsymbol{b}$, where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the multiplicative weights and the additive bias values of the output layer, can represent an approximation to any Borel measurable function (including all continuous functions) with an arbitrary degree of accuracy given that the number of computational nodes at the output of $f^{(1)}(\boldsymbol{x})$ is sufficiently large [49,50,52–55]. However, the universal approximation theorem does not provide any information on how to determine the necessary number of nodes at the output of $f^{(1)}(\boldsymbol{x})$. In fact, it was shown that a shallow network (with one hidden layer) might require exponentially large number of hidden units to accurately represent a function compared to a deeper network model [56]. In addition to this, many examples in the literature empirically demonstrate the advantages of constructing deeper models in terms of the approximation capability and generalization success of the underlying neural network architecture.

When an input $\boldsymbol{x}$ is fed into a network, it goes through a series of transformations denoted by $f^{(i)}$ at each layer $i$, and this process of computing $\boldsymbol{y}$ is called the forward propagation or forward inference. During the training stage, each parameter update step starts with the forward propagation of a batch of inputs. At the final output layer, a cost function, $\mathcal{L}$, with a problem-specific analytical formulation is computed. Deep neural networks are often trained by gradient-based optimization algorithms, e.g., stochastic gradient descent [57–59], that aim to iteratively lower the value of the cost function, $\mathcal{L}$. Gradient-based training with respect to a loss function $\mathcal{L}$ is done through error backpropagation [51]. If we assume, $\mathcal{L} = f^{(2)}(f^{(1)}(\boldsymbol{x}))$, with $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} = f^{(1)}(\boldsymbol{x}) \in \mathbb{R}^n$, then according to the chain rule we have

$$\frac{\partial \mathcal{L}}{\partial x_i} = \sum_j \frac{\partial \mathcal{L}}{\partial y_j}\frac{\partial y_j}{\partial x_i}, \tag{1}$$

which can be written in vector notation as

$$\nabla_x \mathcal{L} = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^T \nabla_y \mathcal{L}, \tag{2}$$

where $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ denotes the $n \times m$ Jacobian matrix of the operator $f^{(1)}$.

In modern deep learning architectures, the functional form of each layer, $f^{(i)}$, is composed of a linear transformation of the output of the previous layer followed by a nonlinear activation function, i.e., $f^{(i)} = \sigma(\boldsymbol{W}^{(i)}f^{(i-1)} + \boldsymbol{b}^{(i)})$, with $\sigma$ denoting the activation function. If $\boldsymbol{W}^{(i)}$ is a full matrix connecting all the neurons on two successive layers to each other, then it is called a fully connected neural network layer. A major part of the computational burden in a forward inference task can be attributed to the linear transformations performed on each layer, consisting of a series of multiplication and summation operations, also known as multiply-accumulate (MAC) operations. This also applies to convolutional neural networks (CNNs), which are especially effective for processing multidimensional visual data. CNNs form a special case, where $\boldsymbol{W}^{(i)}$ is a sparse matrix representing only localized neural connections modeling the convolution operation with learnable weights. In fact, as we discuss in Section 3, the

computational load associated with the MAC operations in feedforward neural networks is one of the main motivations behind the development of various photonic deep learning accelerators and related optical computing techniques based on linear materials.

### 2.2. Historical Overview of Optical Neural Networks and Photonics in Computing

One of the motivations behind optical neural networks in the early days of optical computing stems from the capability of optics in providing high-throughput and high-speed information encoding and processing schemes that can be parallelized. Early generations of optical computing schemes have mostly revolved around the Fourier transform (FT) property of lenses, which is valid under the small-angle approximation [60,61]. One of the most commonly employed optical computing architectures was the $4f$ correlator proposed by different groups [62,63]. The major challenge related to the $4f$ correlators was the accurate implementation of complex-valued Fourier coefficients of a desired spatial filtering function. This limitation was overcome by the seminal work of Lugt [64], giving the $4f$ optical correlator more power, also motivating subsequent research in the field. To realize the optimum matched filter function for any given signal, $s(x, y)$, Lugt recorded the interference pattern created by the FT of $s(x, y)$, $S(p, q)$, and a reference beam over a photosensitive holographic film. The resulting interference pattern, $G(p, q)$ mainly represents three terms:

$$G(p,q) = (|R(p,q)|^2 + |S(p,q)|^2) + R^*(p,q)S(p,q) + R(p,q)S^*(p,q). \tag{3}$$

The first term in Eq. (3) corresponds to a DC component representing the superposition of the FT intensity of the reference, $R(p, q)$ and the signal $S(p, q)$. Although this DC term is of no particular interest, the other two off-axis terms represent the cross correlation operation and its complex conjugate, respectively. The tight optomechanical alignment requirements of the setup in Ref. [64] were relaxed with the help of an off-axis design in [65], termed as the joint transform correlator (JTC). In its original form, JTC is used to optically compare two given real-valued signals using a two-step process. First, these two signals are placed side-by-side, e.g., one at $(x = a, y = 0)$ and the other at $(x = -a, y = 0)$, at the front-focal plane of a Fourier transforming lens and their joint FT spectrum is recorded on a photosensitive film. Once this holographic film is placed at the front-focal plane of another Fourier transforming lens, the output at the back-focal plane will have two separate 2D functions centered around $x = -2a$ and $x = 2a$, corresponding to the cross correlation between the two input signals. Many variants of the original architectures of $4f$ optical correlators and JTC were developed, including nonlinear JTCs [66,67], targeting various applications, including information encryption [68–70], synthetic aperture radar (SAR) [71–74], and pattern recognition [75–78]. Beyond space-invariant convolutions, correlation, and matched filtering operations, coherent optical processing techniques were also adapted to implement a more general family of linear transformations such as the Hough transform [79], matrix–vector multiplications, and coordinate transformations [80].

In addition to these coherent optical computing techniques, optical processing schemes using incoherent light were also proposed. In these systems, the information is encoded in the intensity of the optical field, and this information representation over intensity restricts the range of values to non-negative real numbers. Despite this limitation, incoherent optical processing architectures were developed to realize a series of basic computing tasks, including matrix–vector multiplications [81,82], character recognition [83], FT operation [84], among others [85].

Despite these efforts, these early optical computing techniques have not found widespread practical use, with one exception being the SAR in the early 1960s. On the other hand, the progress on diffractive optical element (DOE) design and fabrication,

optical materials and the spatial light modulator (SLM) technology motivated another surge in optical information processing techniques in the 1980s. A series of adaptive JTC architectures taking advantage of the SLM technology were constructed and used for, e.g., road sign identification [86], and the optical correlator setup was miniaturized to fit into a personal computer [87].

In parallel to these, interest in neural network models also saw a rapid increase due to various important advances in the field [49–51]. Despite the progress in deep learning, however, electronic computers, which could easily outperform the human brain in arithmetic operations, were still falling behind when it comes to complex inference tasks such as pattern recognition. In addition to this, very large-scale integrated (VLSI) circuit technology was facing a challenge due to the increase in the form factor of electronic chips: the computational speed and efficiency of VLSI systems were bottlenecked by wire-based interconnect technology. These limitations motivated extensive research on brain-inspired computing platforms along with faster interconnect technologies for addressing the high computational load in the training of neural networks [88]. Several VLSI circuits were developed for proof-of-concept demonstrations of brain-inspired computing and adaptability [89–92]. In parallel to these efforts, optical neural networks [93–98] were also introduced together with analog optical MAC processors that can all-optically achieve, e.g., matrix operations [99] and systolic array processing [100]. Demonstrating an array of dynamic nonlinear crystals as a planar neural interconnect was another important milestone [101]. This was followed by the seminal work of Li *et al*. [102], in which an optical network that can recognize faces in real-time with very good accuracy was trained by using a photorefractive crystal to store approximately 1 billion weights.

Despite the success of custom neural network hardware demonstrations, the interest in optical neural network technologies slowly faded toward the end of the 1990s. Some of the main reasons behind this were: (1) the high-end GPUs that are cost-effectively used today were not available then, hindering the advances that could be made in both electronic and optical neural networks in terms of both training and inference; (2) the amount of training data were inadequate or hard to acquire in a high-throughput manner to effectively train deep neural network architectures and, when the neural network models are rather small, the motivation for developing analog accelerators is weakened; (3) the optoelectronics and integrated photonics industry and manufacturing techniques were not as mature as today, limiting the capabilities of optical and photonics hardware in terms of performance, scaling, and cost. Over the last two decades in particular, major advances in GPUs, optical materials, micro/nanofabrication technologies, optoelectronics, and integrated photonics have changed the landscape of optoelectronic computing. As a result, optics and photonics fields are now in a much favorable position to prove their potential advantages, particularly on feedforward statistical inference tasks through deep neural network architectures.

## 3. OPTICAL INFERENCE AND COMPUTING

Deep learning [103,104] has become the standard algorithmic tool in processing visual data collected by focal-plane arrays and other optoelectronic sensing technologies. This has fueled the recent revival of research in optical neural networks. The designs of optical networks can be broadly divided into two categories. The first approach focuses on designing optical computing chips that take advantage of integrated photonics, waveguide technology, and related optoelectronic devices. As these optical neural network chips and deep learning accelerators mostly use high-throughput fabrication techniques already heavily used in the electronics industry, they have the potential to provide cost-effective and scalable solutions that are compatible with the current electronic chips for future integration.

As an alternative approach, free-space-based 3D optical neural networks and computing systems have also been demonstrated by exploiting the interaction of light with engineered materials. These optical processors aim to compute a given inference task before the light waves reach the photodetectors or the focal-plane array. From another perspective, these approaches aim to provide free-space propagation-based alternatives to conventional lens-based machine vision systems and can empower task-specific, resource-efficient designs that complete the entire computing task as the light propagates within a material; they can also be jointly trained with back-end electronic neural networks to reduce the computational load, demanding resolution and memory requirements on modern machine vision systems.

In Section 3.1, we first introduce and discuss the recent progress on integrated photonics toward the creation of high-speed neural interconnects and neuro-inspired photonic devices. Section 3.1 ends with the description of photonic neural network architectures that have recently been demonstrated experimentally. Section 3.2 focuses on the second design approach outlined previously, discussing the operational principles and successful demonstrations of free-space-based diffractive optical processors, also introducing hybrid (optical–electronic) network systems.

### 3.1. Integrated Photonics for Statistical Inference and Computing

One of the motivations behind the research on integrated photonics in the 1970s was to develop optical supercomputers, see for example [105]. In those early days, the research was mainly focused on ferroelectric materials, e.g., lithium niobate ($LiNbO_3$), and III–V compound semiconductors such as GaAs and InP. Although $LiNbO_3$ drew attention due to its large electro-optic coefficient enabling light modulation via the Pockels effect [106], III–V semiconductors were also found interesting because of their prospect in laser fabrication, optical amplification, and electronic integration [107].

Over the years, silicon electronics has dominated the industry and, thus, silicon-based fabrication technologies have shown a rapid development, which attracted the attention of optics researchers. That said, silicon photonics had a rough start around the mid-1980s [107,108], due to two major factors. First, the crystalline silicon has an indirect bandgap causing it to be an extremely inefficient material for light emission. Second, crystalline silicon does not exhibit any electro-optic modulation effect due to its centro-symmetrical structure [107]. Despite these difficulties, the research on silicon photonics has grown and matured to play a major role in communication systems as well as in intra- and inter-chip optical interconnect technology, aiming to solve some of the data rate bottleneck and power dissipation issues imposed by electrical wire-based connections [94,109–111]. In fact, the photonic integrated circuit (PIC) technology with silicon-on-insulator (SOI) wafers are already used as the building blocks of modern-day data centers [112].

The success of deep learning and neural networks, as well as the availability of massive amounts of big data in various fields, have brought demanding expectations from computational systems that are hard to meet with the current state-of-the-art electronic hardware. As a result, in recent years, photonics researchers have been focusing on the design of optical communication and interconnect systems for the development of neuromorphic circuits and photonic deep learning accelerators that can be integrated with existing electronic computing systems to enable high-bandwidth, low-latency, and low-power hardware platforms for the future needs of AI applications [3,113].

#### *3.1a. Photonic Neural Interconnects and Deep Learning Accelerators*

The bulk of the computational burden in an inference task achieved through state-of-the-art deep neural networks rests upon the large amounts of MAC operations

corresponding to linear transformations that need to be executed at each layer before the nonlinear activation functions [114]. It is well-known that such linear operations can be efficiently implemented in the optical domain. In the context of PICs, for instance, it has been shown that any $N \times N$ unitary matrix $\boldsymbol{U}$ can be implemented using a mesh of Mach–Zehnder interferometers (MZIs) [115,116]. The theoretical foundation behind this scheme is based on the fact that a matrix $\boldsymbol{U}$ can be constructed by cascading $\frac{N(N-1)}{2}$ rotation operators, $\{\boldsymbol{R_k}\}$, each acting sequentially on 2D subspaces of the full $N$-dimensional vector space [115]. Composed of two beam splitters (or directional couplers) and two controlled phase-delay elements, $\theta$ and $\phi$, each MZI can be used as a four-port device, two input and two output, that can achieve a 2D rotation operator, i.e., $\boldsymbol{R_k}(\theta_k, \phi_k) = \begin{bmatrix} e^{j\phi_k}(e^{j\theta_k} - 1) & je^{j\phi_k}(e^{j\theta_k} + 1) \\ j(e^{j\theta_k} + 1) & 1 - e^{j\theta_k} \end{bmatrix}$.

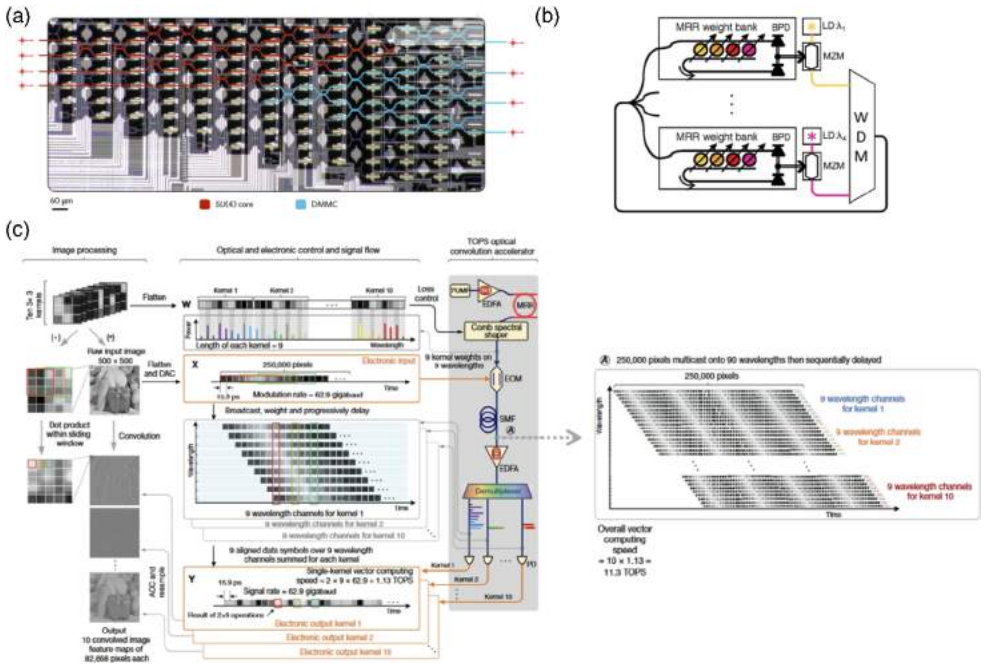
Matrices in neural networks are, in general, not unitary but rather represent arbitrary linear transformations. Fortunately, an MZI mesh system capable of achieving any $N \times N$ unitary matrix could also realize an arbitrary linear transformation matrix, $\boldsymbol{A}$, of size $M \times K$, provided that both $K$ and $M$ are not larger than $N/2$ (see Ref. [117]). In other words, an arbitrary matrix, $\boldsymbol{A}$, can be generated as an $M \times K$ submatrix of an $N \times N$ unitary matrix, if $N$ is sufficiently large, i.e., $K \leq \frac{N}{2}$ and $M \leq \frac{N}{2}$.

Alternatively, an arbitrary $M \times K$ matrix $\boldsymbol{A}$ can also be physically implemented through singular value decomposition (SVD), $\boldsymbol{A} = U\Sigma V^{\dagger}$, provided that optical attenuators and/or amplifiers are embedded into the optical path on the photonic chip [118], as shown Fig. 1(a). Although the unitary matrices, $U$ and $V$, are already suitable for MZI-based implementations, the diagonal matrix $\Sigma$ requires the integration of optical attenuators and/or amplifiers in the form of, e.g., semiconductors [119], dyes [120], and phase change materials (PCMs) [121]. For instance, a successful realization of arbitrary weight matrices of a fully connected neural network was experimentally demonstrated by cascading a few MZI mesh architecture with embedded attenuation paths [118].

An important challenge of these MZI-based optical processors is that the number of interferometers and integrated components in the photonic chip scales exponentially with the size of the desired transformation matrices. For instance, implementing a $N \times N$ unitary linear transformation requires $\frac{N(N-1)}{2}$ MZIs or, in other words, a total of $N(N-1)$ beam splitters/couplers combined with $N(N-1)$ phase shifters; including the $N$ additional phase shifters at the output ports, the total number of phase shifters that are needed becomes $N^2$. If the $N \times N$ matrix is not unitary, then these numbers further increase, significantly hindering the scalability of these systems toward processing, e.g., high-resolution images with several million pixels. Moreover, power losses, noise, and other physical imperfections increase proportionally with the number of integrated components in the system. Proposed methods toward mitigating such unwanted system defects include, e.g., fault-tolerant MZI mesh architectures [117,124], designs that incorporate component imperfections [125], "self-configured" MZIs [126,127], and *in-situ* training of neuron connection weights [128].

Tuning the connection weights and phase delays inside the silicon waveguides with a fast, power-efficient, and robust mechanism represents yet another key challenging aspect of integrated photonic computing. Some of the most notable effects that can be used for adjusting the refractive index of a region within silicon waveguides include the thermo-optic effect, free-carrier absorption, and free-carrier dispersion (plasma dispersion) [113]. Although thermal tuning mechanisms based on metal filament microheaters [129] and waveguide-embedded heaters [130] are among the most robust and convenient methods, they are slow and power-inefficient. Instead, for faster operation, one can fabricate hybrid waveguides made of a silicon core surrounded and/or

Figure 1




**Photonic interconnects and deep learning accelerators.** (a) MZI mesh that can all-optically realize $4 \times 4$ unitary transformation (red pathway) and attenuation (blue pathway) [118]. Reprinted by permission from Macmillan Publishers Ltd: Shen *et al.*, Nat. Photon. **11**, 441 (2017) [118]. Copyright 2017. (b) Concept of a broadcast-and-weight network with modulators used as neurons. MRR, microring resonator; BPD, balanced photodiode; LD, laser diode; MZM, Mach–Zehnder modulator; WDM, wavelength-division multiplexer [122]. Reprinted by permission from Macmillan Publishers Ltd: Tait *et al.*, Sci. Rep. **7**, 7430 (2017) [122]. Copyright 2017. (c) Photonic implementation of a convolutional layer with 10 $3 \times 3$ filter channels processing an input image of size $500 \times 500$ based on the time- and wavelength-multiplex deep learning accelerator presented in Ref. [123]. This large-scale, photonic deep learning accelerator can achieve 11 TeraOps/s. Reprinted by permission from Macmillan Publishers Ltd: Xu *et al.*, Nature **589**, 44 (2021) [123]. Copyright 2021.

doped with other materials, such as III–V semiconductors [131], $LiNbO_3$ [132], and graphene [133], to tailor the refractive index modulation. However, these methods can provide only a limited dynamic range for optical modulation and they create photonic structures that are susceptible to electrical damage [113]. As an alternative solution, PCM-based all-optical methods [121,134,135], which do not require any external electrical or thermal input for tuning the material properties, have been proposed. Instead, these all-optical neural interconnects exploit optically induced changes in, e.g., $Si_3N_4$ [121], metal sulfide fibers [134,135] to control the light propagation inside silicon waveguides. An additional benefit related to all-optical PCM-based solutions is that these materials are non-volatile, i.e., they do not require any external source to maintain their state. Hence, they offer a power-efficient and fast way to adapt the linear weights along silicon waveguides. Therefore, they could potentially play a major role in the future of photonic deep learning accelerators and the associated neural network architectures.

In addition to the spatial modes exploited by the MZI-mesh circuits in realizing weighted interconnects, the optical waves have other orthogonal features, e.g., wavelength and polarization, that do not interact with each other and, thus can

independently be used to encode and/or carry information. There have been several works that utilize the spectral composition of optical waves to physically realize MAC operations in photonic circuits [121,122,136–139]. According to the proposed scheme in [136], a synaptic connection is realized through a bank of microring resonators (MRRs) that acts as tunable filters to process a series of spectral subbands in parallel, termed as broadcast-and-weight (see Fig. 1(b)). Originally proposed as a photonic MAC architecture for neuromorphic spiking neural networks [140], it also inspired several other deep learning accelerator designs based on optical microdisk arrays [141] and memristor-driven MRR banks [142,143]. A variant of the broadcast-and-weight architecture that is optimized specifically for the implementation of photonic convolutional layers was proposed in Ref. [144], where the authors exploited the localized and sparse nature of connections in convolutional layers to reduce the number of MRRs, resulting in a more compact system. The broadband photonic neural interconnect designs demonstrated in [121,138] are also similar to the broadcast-and-weight scheme in the sense that they process the weights of the spectral subbands simultaneously. However, they require an additional demultiplexing step to achieve this parallelism, as the PCMs [121] and InP semiconductor optical amplifiers (SOAs) [138] lack the spectral selectivity of MRRs.

Even at this relatively early stage of photonic AI accelerators, there has been crucial and promising progress towards the scalability of MAC processors. Recently, Xu *et al.* [123] reported a photonic accelerator for CNNs that can realize a convolutional layer with 10, 3×3 filters acting on input images of size 500×500 pixels as illustrated in Fig. 1(c). To achieve this, they multiplexed the optical waves both in time and spectral domains (see Fig. 1(c)). In this scheme, the flattened/vectorized weight vector representing the coefficients of a convolutional filter is encoded into the power spectrum of the light waves. Based on this encoding, the optical power of the $i$th spectral component is set to be $\boldsymbol{W}[N-i+1]$ with $i \in [1, N]$, where $N$ denotes the size of the flattened convolutional kernel (for example, $N = 9$, for a 3×3 filter). Meanwhile, the input image is also vectorized into a 1D vector, $\boldsymbol{X}$, of size $L$ (for example, $L = 25 \times 10^4$, for a 500×500 image) and represented as the amplitude of a stepwise electrical waveform, $\boldsymbol{X}[n]$ with $n \in [1, N+L-1]$ including zero padding. An electro-optic Mach–Zehnder modulator is driven by this input electrical waveform, $\boldsymbol{X}[n]$, to broadcast the input gray levels values onto the shaped optical comb lines. As a result, the optical power at the $i$th wavelength channel becomes $\boldsymbol{W}[N-i+1]\boldsymbol{X}[n]$. The modulated optical signal is, then, progressively shifted in the time domain with the help of a dispersive fiber that applies a wavelength-sensitive time delay equal to the duration of a step of the input electrical waveform $\boldsymbol{X}[n]$. Consequently, the optical power of the shifted replica at the $i$th spectral component becomes $\boldsymbol{W}[N-i+1]\boldsymbol{X}[\mathrm{n-i}]$. Setting the duration of integration at the photodetector to be equal to the time delay between each spectral component, the total spectral power at every time point, $n$, is accumulated, creating the output electrical waveform, $\boldsymbol{Y}[n]$:

$$\boldsymbol{Y}[n] = \sum_{i=1}^{N} \boldsymbol{W}[N-i+1]\boldsymbol{X}[n-i], \tag{4}$$

which corresponds to the discrete convolution between the input $\boldsymbol{X}$ and the filter $\boldsymbol{W}$. At the readout of the photodetector, each sample of $\boldsymbol{Y}[n]$ within the range $n \in [N+1, L+1]$, corresponds to the inner product between $\boldsymbol{W}$ and a region of input image $\boldsymbol{X}$. This photonic computing framework can also be extended to implement any linear transformation by simply adopting the sampling intervals at the readout of the photodetection [123]. Figure 1(c) depicts a case where $\boldsymbol{W}$ describes a convolutional network layer with multiple channels, then the weights of each 2D filter are encoded into the optical power of wavelength components within a distinct spectral subband.

Therefore, the required optical bandwidth of operation is directly proportional to the size (width, height, and depth) of the targeted convolutional layer to be implemented. The authors reported 11.3 TeraOps/s by utilizing only a 36-nm bandwidth covering mostly the C-band (1540–1570 nm). Although this performance is inferior compared with Google TPUs (tensor processing units) and other chips offering >200 TeraOps/s, the authors provide possible improvement directions, e.g., utilizing the entire telecommunication band (1460–1620 nm). In addition, if the presented time and wavelength multiplexing can be further enriched with the incorporation of polarization and spatial modes, photonic deep learning accelerators might enable PetaOps/s operation for CNNs with more than 24,000 synaptic connections [123].

#### *3.1b. Neuromorphic Computing Using Photonics*

Despite the remarkable technological advances in nanofabrication technologies and VLSI circuit architectures, conventional electronic computer systems are far from competing with the performance and efficiency of the human brain. The large gap between electronic computers and the brain can be attributed to several important factors. Although the traditional computers are centralized, digital devices, the brain represents information encoded in the relative values of analog signals, and it computes in a decentralized, highly parallel manner [88,92]. Neuromorphic engineering is broadly concerned with the development of physical hardware systems that can potentially mimic the neuro-biological structure and fundamental operational principles of the nervous system. In relation to that, neuromorphic computing aims to bring the efficacy of biocomputing into engineered computational devices [89,92], and it remains to be an active area of research both in electronics [145–149] and photonics [150–152].

Towards realizing brain-inspired optical computing hardware, one of the key challenges faced by neuromorphic photonics has been to implement activation functions analogous to the nonlinear action potential dynamics of biological neurons [113,153,154]. In fact, the difficulty of realizing nonlinear activation functions at acceptable optical intensity levels in a power-efficient and scalable manner could be considered as one of the reasons behind the gradually vanishing interest in optical neural networks during the 1990s [3]. During the past two decades, on the other hand, there have been numerous advances in the fabrication of silicon photonic chips and related components, enabling researchers to revisit neuro-inspired photonic devices with nonlinear physical dynamics. Some of these recent research efforts toward realizing photonic nonlinearities can be divided into two mainstream approaches based on the physical domain of signal flow between the input and output ports of these devices: (1) optical-electrical-optical (O/E/O) and (2) all-optical [113,150].

The O/E/O type of neuron-like photonic devices offer a degree of design flexibility because the input and output ports are optically isolated, meaning that the properties of the output optical waves, e.g., wavelength, power, can be set independently from the properties of the input waves. In addition, the electrical domain in the signal pathway can be exploited to implement, enhance, and/or tune the nonlinearity of the device. In the O/E/O approach, the first O/E conversion step is, in general, responsible for the accumulation/summation of the weighted signals at the input of the nonlinearity and is often implemented by a photodetector or array of detectors. As the photodetector integrates the incoming photons, an electric current that is proportional to the total intensity of the collected optical waves is generated. An important distinction among different O/E/O approaches proposed in the literature can be attributed to the mechanism of establishing the nonlinear response. Although some of the O/E/O devices apply the nonlinearity purely in the electrical domain through, e.g., superconductors [155,156] (see Fig. 2(a)) or digital lookup circuits [118,138], others utilize the E/O conversion

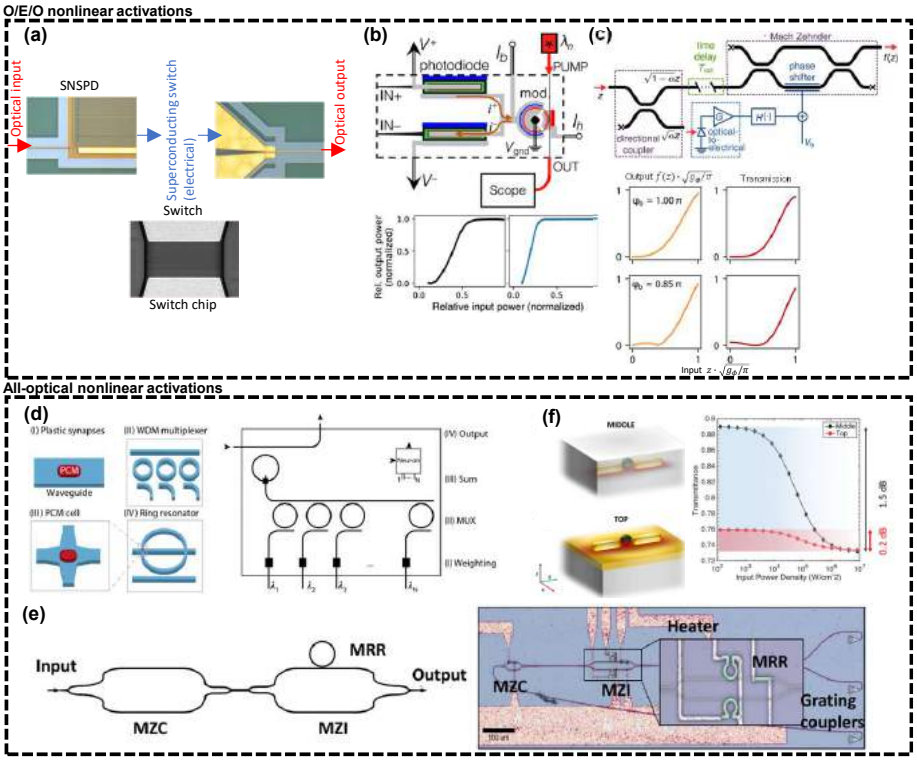
stage designed around lasers [157] and/or modulators [158–160] (see Fig. 2(b)). A MZI-mesh compatible nonlinear O/E/O node design has recently been reported by Williamson *et al*. [161], in which the optical signal at the input is split into two arms, with one arm accommodating a photodetector followed by an electronic phase shift controller, determining the nonlinear transfer function at the intersection output of the two arms as a function of the input as shown in Fig. 2(c). In addition to its MZI-mesh compatible structure, this approach also allows the nonlinear transfer function to be tuned by changing the bias voltage over the phase shift controller arm. The authors demonstrated ReLU-like functions along with clipped nonlinear responses akin to bounded activations. A similar tunability is also reported in Ref. [158] (see Fig. 2(b)), where the authors generated sigmoid and ReLU-type functions along with a radial basis function highlighting their design flexibility.

Unlike the O/E/O approaches, the information is not represented in the electrical domain inside an all-optical photonic neuron. Instead of using E/O conversion dynamics and/or electrically controlled modulators, the nonlinear activation functions are achieved based on manipulating the material properties, such as optical susceptibility and carrier concentrations in semiconductors, triggered by light–matter interactions. Inducing changes in optical susceptibility of nonlinear materials is, in general, power inefficient, meaning that the output of the photonic neuron is much weaker than its input. In addition, cascading these photonic structures becomes challenging and often results in extremely photon inefficient designs. These issues, however, can be mitigated by using a form of carrier regeneration/injection as first shown by Hill *et al*. almost two decades ago [163], and this approach has been used in experimental demonstrations of all-optical photonic neurons. These all-optical photonic neuron designs include carrier regeneration over light–semiconductor interaction principles: cross-gain modulation [164] and cross-phase modulation [165]. One of the challenges associated with these carrier regeneration techniques is to isolate the controlling input signal from the output signal. To partially address this challenge, the device parameters are often tuned to generate weaker output signals than the input, and they are used in conjunction with optical amplifiers that boost the signal to an adequate level to drive the neurons on subsequent layers.

Other promising approaches toward realizing all-optical nonlinear activation functions have also been presented based on PCMs [121,166], MRR-loaded arms in MZI architectures [167], and nanoparticles embedded inside silicon waveguides [162], see, e.g., Figs. 2(d)–2(f). In Ref. [121], the authors integrated PCM cells as part of MRRs to induce switching behavior based on the PCM state, amorphous or crystalline, which depends on the incident optical power over the material. Specifically, with the PCM in its crystalline state, the optical injection signal and the MRR is in resonance with each other. Consequently, all the injected signal is coupled to the MRR before reaching the output node. When the intensity of the light incident on the PCM material surpasses a certain threshold, the material goes to an amorphous state, disrupting the resonance between the MRR and the injected optical signal. In this state, the optical pump signal bypasses the MRR and directly reaches the output node, resulting in a transfer function that resembles ReLU.

Among the photonic neuron design approaches discussed thus far, the O/E/O neurons, where the information is digitized in the electrical domain [118,138] to generate the nonlinear response, operate at the slowest rates due to the bottleneck created by the analog-to-digital converters (ADCs) in the signal pathway. On the other hand, in all-analog O/E/O neuron designs without any digitization, the primary factor limiting the speed of operation is, in general, not the O/E or E/O conversions, but rather the relatively slow carrier drift in semiconductors and/or electrical wire-based connections

Figure 2



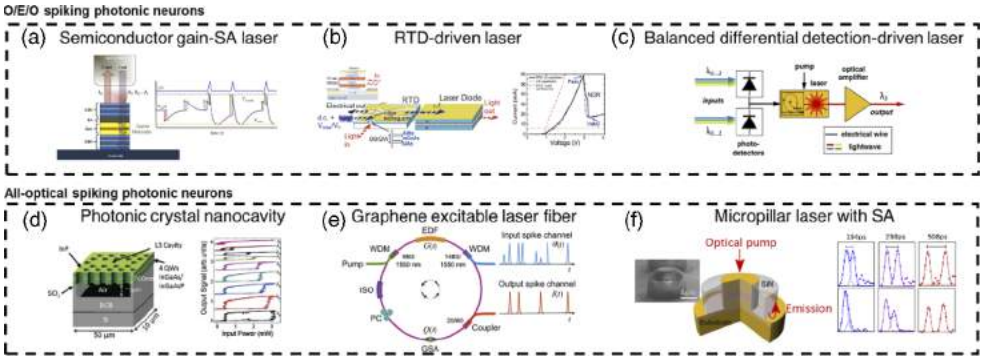

**Photonic implementations of nonlinear activation functions.** (a) A superconducting optoelectronic neural activation device [156]. A superconducting-nanowire single-photon detector (SNSPD) converts the incident light into electrical current and drives a superconducting switch followed by an integrated LED that emits the output light. Reprinted by permission from Macmillan Publishers Ltd: McCaughan *et al.*, Nature Electron. **2**, 451 2019 [156]. Copyright 2019. (b) Silicon photonic modulator neuron, as proposed in Ref. [158]. The input optical intensity is converted into electrical current driving a ring modulator to exploit its nonlinear response. The nonlinear input/output response of the device is controllable: see the two exemplary response curves (bottom). Figures 2 and 5 reprinted with permission from Tait *et al.*, Phy. Rev. Appl. **11**, 064043 2019 [158]. Copyright 2019 by the American Physical Society. (c) An MZI-based O/E/O-type nonlinear activation unit [161], where a photodetector drives an adjustable phase shifter controlling the interference on two arms. The nonlinear response can be controlled by changing the bias of the phase shifter. © 2019 IEEE. Reprinted, with permission, from Williamson *et al.*, IEEE J. Sel. Topics Quantum Electron. **26**, 1–12 (2019) [161]. (d)–(f) All-optical photonic nonlinear activation device models. (d) Fundamental building blocks of an all-optical neural network that uses PCMs to implement both synaptic connections and neural activations [121]. The weighted neural connections are achieved by WDM multiplexer–demultiplexers (II) and PCMs embedded in silicon waveguides (I). The nonlinear activation is achieved by embedding a PCM cell (III) into the input junction of a ring resonator (IV). Reprinted by permission from Macmillan Publishers Ltd: Feldmann *et al.*, Nature **569**, 208 (2019) [121]. Copyright 2019. (e) Schematic of an MRR-loaded MZI-based nonlinear thresholder. On the right, micrograph of the all-optical activation device. (f) A gold nanoparticle (NP) and CdSe quantum dot (QD) embedded in the middle or on the top of a silicon waveguide implement an all-optical nonlinear response, red (black) curve depicting the activation when the NP/QD is placed on top (in the middle) of the waveguide. Reprinted with permission from [162]. Copyright 2018 Optical Society of America.

[113]. The time delay and power consumption in these O/E/O systems are mostly related to resistance–capacitance constants of different individual parts constituting the device. Although there are O/E/O designs exhibiting very small intrinsic capacitances, e.g., ~2 fF as reported in, e.g., Ref. [168], the all-optical photonic neurons still maintain their status as being the fastest. One drawback related to the all-optical photonic neurons is that they lack the tunability offered by the O/E/O approaches, meaning that the functional forms of the realizable nonlinear activation functions are constrained by the response of the underlying light-matter interactions. Nonetheless, the recent studies on the universal approximation theorem suggests that the class of nonlinear functions that ensure theoretical guarantees for the inference capabilities of deep neural networks, might be larger than previously predicted [53,54]. Furthermore, several works on photonic neural networks have shown that the data-driven training techniques of deep learning can be adapted to physically attainable nonlinear transfer functions without major performance degradation [118,122,169].

Beyond the continuous-time photonic nonlinearity models outlined previously, there has been extensive research on photonic systems that aim to approximate the spiking dynamics of the nervous system [170–172] using spiking photonic devices as their computational primitives [121,140,173–189]. Unlike the transfer function of continuous-time nonlinearities, which can be described by a first-order differential equation, $\dot{y} = h(y, x)$, where $x$ and $y$ denote the input and output signals, respectively, the operation principles of spiking devices rely on time-varying internal state variables, $z$, and there is not one but rather a set of differential equations describing their operation; $\dot{z} = f(z, x)$ and $\dot{y} = g(y, x, z)$ [173,182]. The operational dynamics of these devices simply have three main behavioral regimes. In the first regime, the system rests in an equilibrium without any perturbation. When a perturbation above a certain threshold is applied, the system deviates from the equilibrium conditions triggering a burst of optical power in the form of a single pulse or a series of pulses. The perturbation level that is required for the transition from the equilibrium regime to pulsing mode is called the excitability threshold. Following the burst of optical power, the system returns to its equilibrium state, and the duration of this settlement regime is called refractory period which directly determines the pulse firing rate of the device.

One of the most commonly used types of integrated photonic devices aiming to mimic the biophysical dynamics of neurons is based on excitable semiconductor lasers [173]. Widely studied semiconductor-based optical excitability models include; two-section gain and saturable absorber (SA) lasers [177,178] (Fig. 3(a)), semiconductor ring [180] and micro-disk [183] lasers, photonic crystal nanocavities [181] (Fig. 3(d)), injection-locked semiconductor lasers [188], semiconductor lasers with optical feedback mechanisms [189], micropillar lasers with embedded SAs [184,185] (Fig. 3(f)), polarization switching in vertical-cavity surface-emitting lasers (VCSELs) [174], graphene excitable lasers [186,187] (Fig. 3(e)), and resonant-tunneling diode (RTD)-driven lasers [190] (Fig. 3(b)). Broadly speaking, these excitable photonics designs have a gain medium, a saturable process, and a cavity, all in a single device. The pump of the gain medium can either be an electrical or an optical signal, whereas the injection modes can broadly be classified as: (1) coherent optical injection, (2) incoherent optical injection, and (3) electrical injection. In a coherent injection device, the input and output signals occupy the same spectral band and they offer compatibility with coherent optical processing architectures. However, they require a precise global phase control which is challenging to achieve, particularly in systems of synchronized lasers. For incoherent injection devices, the input and output optical signals are at different wavelengths. In some cases, the input signal can also be used as the pump provided that the wavelength of the input signal is smaller than the wavelength of the output [186,187].

Figure 3




**Excitable lasers and cavities for photonic spiking.** (a) Two-section gain medium that integrates the input power combined with a saturable absorber excitable laser acts as an integrate-and-fire neuron [177]. Right, an illustration of spiking dynamics of the device. When enough excitatory input current arrives at the input port, it drives the voltage $V(t)$ (purple) above a threshold unleashing a spike $y(t)$ at the output (blue). © 2013 IEEE. Reprinted, with permission, from Nahmias *et al*., IEEE J. Sel. Topics Quantum Electron. **19**, 1–12 (2013) [177]. (b) An O/E/O neuron that is composed of resonant-tunneling diode (RTD) layer stack, photodetector, and laser diode, exhibiting spiking behavior [190]. Right, excitability is achieved by biasing a double-barrier quantum well (DBQW) within the RTD in the negative differential resistance (NDR) region of within its current–voltage curve. Reprinted with permission from [190]. Copyright 2013 Optical Society of America. (c) A spiking photonic neuron that operates based on a pair of balanced differential photodetector pair driving a laser followed by an optical amplifier [140]. This device uses incoherent injection, meaning that the wavelength of the light at the output is different than the wavelengths at the input. (d) InP-based 2D photonic crystal nanocavity with quantum wells (QWs) [181]. Right, the hysteresis demonstrates the bistable operation depending on the cavity resonance. Figures 1 and 2 reprinted with permission from Brunstein *et al*., Phys. Rev. A **85**, 031803 (2012) [181]. Copyright 2012 by the American Physical Society. (e) Graphene-SA excitable laser fiber, as proposed in Ref. [186]. An erbium-doped fiber is optically pumped and used as a gain medium. Reprinted by permission from Macmillan Publishers Ltd: Shastri *et al*., Sci. Rep. **6**, 19126 (2016) [186]. Copyright 2016. (f) An optically pumped III–V semiconductor micropillar with SA acting as a spiking photonic device. Right, recorded time traces of input perturbations (upper), and the system responses (lower), when the bias pump is set to be 71% of self-pulsing threshold. Figures 1 and 3 reprinted with permission from Selmi *et al*., Phys. Rev. Lett. **112**, 183902 (2014) [184]. Copyright 2014 by the American Physical Society.

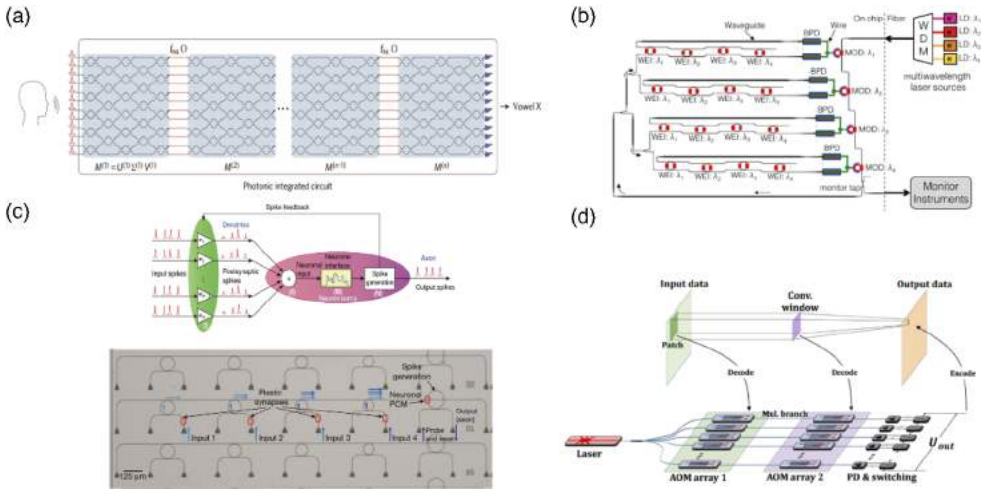
While the progress on the design of photonic spiking devices has been rapidly developing, the experimental studies investigating their cascadability and compatibility with the existing weighted neural interconnect technologies are still limited. Recently, Nahmias *et al*. [140] have fabricated a photonic chip with multiple spiking neurons to test their cascadability (see Fig. 3(c) for the proposed nonlinear activation device). Their neuron-like spiking device model, which comprises a balanced photodetector pair, a two-section distributed feedback lasers, and a SOA operating based on incoherent injection, is also compatible with the broadcast and weight scheme presented in Ref. [136]. Robertson *et al*. [176] experimentally demonstrated VCSEL-based spiking neurons collectively working to achieve functional processing tasks such as coincidence detection and pattern recognition by processing ultrafast input signals composed of ~100 ps pulses. In another effort, Feldman *et al*. [121] experimentally demonstrated a

single-layer, all-optical spiking network with four neurons that can successfully classify four different, binary, 3×5 input images (see Fig. 4(c)). Each neuron has 15 (3×5) synaptic connections weighted all-optically via non-volatile PCM materials, which creates a high transmission contrast between amorphous and crystalline states (see Fig. 2(d)). Instead of semiconductor-based solutions, their all-optical neurons also use PCM cells embedded inside MRRs together with an incoherent optical injection to drive the spiking dynamics as demonstrated in Fig. 2(d). In their experimental system, the value of each pixel of a given input pattern is encoded into the amplitude of a wavelength channel. These 15 wavelength channels are replicated 4 times and distributed over 60 (3×5×4) waveguides using a wavelength-division-multiplexing (WDM) scheme based on MRRs. The input values are then weighted by using PCM embedded waveguides corresponding to the matrix of a 15×4 fully connected network. The signal inside 15 waveguides reaching to each neuron is then accumulated by an additional MRR based WDM system and directed onto the PCM cell of the corresponding neuron, which applies a ReLU-like nonlinearity.

One challenge in photonic spiking neural networks stems from the inefficiency of simulating these continuous-time, dynamic systems with digital clock-based electronic computers. This is also the primary reason behind the stronger prominence of other machine learning modalities. Despite the analogy between the spiking networks and the biophysical dynamics of the nervous system [170–172] and the established mathematical, code-theoretic foundation [192–194] supporting the efficacy of spike-based elasticity in information encoding and processing [195–204], these systems have, in general, been partially shadowed by other neural network schemes more compatible with the digital computing. Future advances in spiking photonic neural network systems might, therefore, be revolutionary in the field of neuromorphic engineering by filling a need for fast, analog hardware platforms to study spiking dynamics of the brain.

Neuromorphic photonic hardware platforms, on the other hand, are also concerned with the optical implementations of more conventional and widely used neural network architectures. For instance, Ref. [118] experimentally realized a feedforward, fully connected network using externally trained MZI-meshes and digital O/E/O nonlinearities following the theoretical model of saturable absorbers. Their system was shown to achieve 76.7% accuracy for the task of vowel classification, as shown in Fig. 4(a). Similarly, a photonic chip based on an MZI-mesh that can physically implement 6×6 arbitrary complex-valued matrices has been fabricated in [205], where the authors compared the real- and complex-valued neural networks for (1) the implementation of an XOR gate, (2) classification of the Iris dataset, (3) binary categorization of spiral and circle patterns, and (4) the MNIST image dataset. With a 4-layer complex-valued fully-connected network with 784, 4, 4, and 10 neurons, respectively, they reported 86.5% accuracy, computed over 200 blind testing samples, for the recognition of MNIST handwritten digits. This 86.5% accuracy decreases to 82.0% when the same network architecture is implemented using only real-valued weight matrices highlighting the benefits of the complex-valued neural networks, which has recently attracted more attention in the deep learning community [206–208]. On the other hand, there are two drawbacks related to this implementation. First, they implemented nonlinear activations purely in the digital domain following signal detection with photodetectors, which limits the computational speed. In addition, due to the limited size of their photonic chip, the weighted connections between the 1st and 2nd layers ($784 \times 4$) as well as the weights of the output layer ($4 \times 10$) were implemented electronically [205]. Recent proof-of-concept experimental demonstrations also include a photonic recurrent [122] and a feedforward neural network [158] (see Fig. 4(c)) designed around the

Figure 4




**Photonic neural network implementations.** (a) A fully connected photonic neural network operates based on MZI meshes to classify vowels [118] (see also Fig. 1(a)). The nonlinear activations are implemented digitally following an analytical model describing the transfer function of saturable absorbers. Reprinted by permission from Macmillan Publishers Ltd: Shen *et al.*, Nat. Photon. **11**, 441 2017 [118]. Copyright 2017. (b) A fully connected feedforward network architecture realized by the broadcast-and-weight scheme [136] (as in Fig. 1(b)) that uses microring resonator (MRR) weight banks depicted here as WEI for photonic neural connections. Each balanced photodetector pair (BPD) corresponds to the input port of a neuron, thus, there are four neurons in total shown here. The BPDs integrate the input signal corresponding to summation of the inner product and convert the information into electrical current to drive an MRR-based modulator to realize nonlinear activation [158]. Figure 1 reprinted with permission from Tait *et al.*, Phys. Rev. Appl. **11**, 064043 (2019) [158]. Copyright 2019 by the American Physical Society. (c) An all-optical spiking neural network implementation using the PCM-based synaptic connection and nonlinear activation schemes shown in Fig. 2(d) [121]. The network has 4 neurons and 15×4 neurosynaptic connections with self-learning capability. Reprinted by permission from Macmillan Publishers Ltd: Feldmann *et al.*, Nature **569**, 208 (2019) [121]. Copyright 2019. (d) A proposed optical convolution unit [191] where the gray levels of the input image pixels are mapped to voltages driving the first acousto-optic modulator array (AOM array 1) and the values of convolutional kernels drive the AOM array 2. Reprinted with permission from [191]. Copyright 2019 Optical Society of America.

broadcast-and-weight scheme reported in Ref. [136], as well as a CNN architecture physically realized through acousto-optic modulators [191] (Fig. 4(d)).

### *3.1c. Reservoir Computing based on Guided Waves and Integrated Optics*

One of the first implementations of reservoir computing was conceived by Verstraeten *et al.* [209], in an effort to unify the underlying principles of echo state networks [210] and liquid state machines [211], developed by Jaeger and Haas and Maass *et al.*, respectively. One of the primary purposes behind these developments was to circumvent the difficulties in training deep recurrent neural networks (RNNs), which offer an attractive deep learning model to process time-series data. Typical reservoir computing systems consist of three major components: one input layer, the reservoir, and an output layer. The $k$-dimensional input information is injected into a reservoir with $N$-nodes through the input connection matrix, $\boldsymbol{W}_{in}$ of size $k \times N$ corresponding

to the input layer. The reservoir, on the other hand, is represented as a dynamical system where the connectivity is modeled through a matrix, $\boldsymbol{W}_{res}$, of size $N \times N$ with $N$ denoting the number of nodes within the system. Earlier work used randomly distributed weights for the input, $\boldsymbol{W}_{in}$, and the internal connections, $\boldsymbol{W}_{res}$. In other words, the training solely optimizes the connections between the nodes of the reservoir and the output, i.e., $\boldsymbol{W}_{out}$, significantly easing the learning process.

In the discrete-time domain, the forward model of a reservoir computing system can be described as

$$\boldsymbol{x}[n] = f(\boldsymbol{W}_{in}\boldsymbol{x}_{in} + \boldsymbol{W}_{res}\boldsymbol{x}[n-1] + \boldsymbol{b}) \tag{5a}$$

$$\boldsymbol{x}_{out} = \boldsymbol{W}_{out}\boldsymbol{x}[n], \tag{5b}$$

where the vectors $\boldsymbol{x}[n]$ , $\boldsymbol{x}_{in}$, and $\boldsymbol{x}_{out}$ denote the internal state of the reservoir at time instant $n$, the input information and the corresponding output, respectively. The term $\boldsymbol{b}$ in Eq. (5) represents a vector of additional biases, whereas the function $f$ is a nonlinear node-wise activation function. According to the inference model described by Eq. (5), the internal state of the reservoir is represented as a nonlinear function of the input information and the previous state of the system, which, in return, points to a fading memory. Moreover, from Eq. (5b), the result of an inference task, $\boldsymbol{x}_{out}$, generated as a response to an input, $\boldsymbol{x}_{in}$, is simply a linear combination of the reservoir state, $\boldsymbol{x}[n]$. Therefore, the reservoir can be considered a nonlinear dynamic system of filters that transforms the data into a higher-dimensional space where the information can be linearly interpretable.

Although the stochastic gradient descent-based, iterative optimization algorithms can be used to train the output connections of a reservoir computing system, the common practice is to compute the $\boldsymbol{W}_{out}$ in a single step using ridge regression based on the following algebraic form:

$$\boldsymbol{W}_{out} = \boldsymbol{M}_{out}\boldsymbol{M}_{in}^{T}[\boldsymbol{M}_{in}\boldsymbol{M}_{in}^{T} + \delta\boldsymbol{I}]^{-1}. \tag{6}$$

In Eq. (6), $\boldsymbol{M}_{in}$ is a matrix that contains the concatenated internal states of the reservoir, $\boldsymbol{x}[n]$, for a series of different input vectors from the target dataset and $\boldsymbol{M}_{out}$ is the matrix of the desired outcomes for a given computational task and the set of inputs in $\boldsymbol{M}_{in}$ . The multiplicative factor $\delta \ll 1$ is a small regularization term that provides robustness against potential ill-posedness and large condition number of the matrix $\boldsymbol{M}_{in}\boldsymbol{M}_{in}^{T}$, which needs to be inverted. Computation of the optimal output connections using a single-step regression process as depicted in Eq. (6), instead of iterative and computationally time-consuming gradient-descent-based approaches, enables the reservoir computers to be trained faster.

Various physical reservoir computing hardware has been developed based on, e.g., water ripples [212], tensegrity structures [213], soft bodies [214], and photonics devices [215–217]. The engineering fields have embraced the concept of reservoir computing due to its advantageous features regarding its implementation and robustness. First, in theory, any physical system with a sufficiently high-dimensional phase space can serve as a reservoir [216,218]. Although it depends on the input–output data spaces and the desired computational task, typical examples of reservoirs operate based on several hundreds of nodes, which can readily be found in most physical systems. Second, the forward model within the reservoir does not need to be optimized, meaning that the physical connection weights within the computing medium do not need to be adjustable. In fact, there are reservoir computing platforms, where the connectivity matrix, $\boldsymbol{W}_{res}$, and the nonlinear activation $f$ can only be modeled partially and the characterization of the reservoir response is done purely based on empirical input–output testing [219]. Furthermore, it has been shown that the connectivity

within the reservoir does not have to be random in nature and can exhibit a certain type of brain-inspired topology [220]. Even more simplistic, non-random approaches such as cyclic reservoirs [221], have also been shown to offer promising computational performance.

Although the nature and complexity of the internal connections do not pose strong prerequisite conditions, a physical nonlinear dynamic system must generate the same output in response to the same input, i.e., it should ideally define a one-to-one nonlinear mapping between the input and output vector spaces. However, satisfying this condition in a dynamic system can be particularly challenging as it requires controlling the initial conditions and the timing of the input injection in an accurate manner. In the case of using natural and physical engineering systems as computing reservoirs, it could be very difficult to monitor and alter the internal state of the reservoir. Thus, a nonlinear physical system is generally required to show the echo state property (ESP) [222,223] to serve as a reservoir, meaning that its internal state only depends on the sequence of inputs previously injected into the system. Otherwise, the state of the reservoir must be controllable by altering external parameters, e.g., electrical signals controlling the transmission coefficients over the pixels of a SLM [224,225].

In their pioneering work, Vandoorne *et al*. presented a 16-neuron reservoir based on passive components such as optical splitter/combiners and waveguides fabricated on a SOI platform [226]. The neurons were placed over a $4 \times 4$ grid. Coupling and splitting at each neuron were achieved by $1 \times 2$ or $2 \times 2$ multimode interferometers. Among all the 16 neurons, one neuron was used to inject input signals, and the intensity levels of 11 neurons were measured by photodiodes as the output of the reservoir. Each neuron was connected to its adjacent neighbors through a 2-cm waveguide. The purpose of the relatively long waveguide was to slow down the photonic reservoir's time scale for additional flexibility in optimizing the ratio of interconnection delay to bit-rate, which was scanned between 125 Mbit $s^{-1}$ and 1.25 Gbit $s^{-1}$. The reservoir was designed fully passive and did not have any nonlinearity other than the intensity detection at the output, which creates signals proportional to the magnitude-squared of the amplitude of a complex electric field, $\boldsymbol{I} \propto |\boldsymbol{E}|^2$. They have shown that a 16-node reservoir computing system operating based on the nonlinear optoelectronic conversion at the photodetectors is sufficient to perform nonlinear logical operations, e.g., XOR, header recognition and classification of spoken digits.

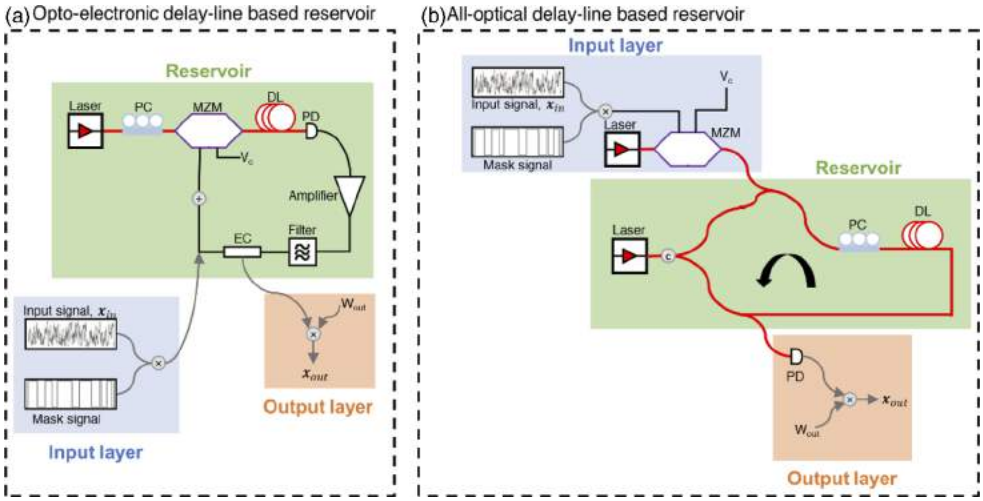
A possible limitation of the photonic reservoir presented in Ref. [226] is related to the node topology, called the swirl architecture [215]. In this node topology, the neuron connections are non-symmetrical, i.e., some neurons use $1 \times 2$ splitter/combiners, and others use $2 \times 2$ optical splitter/combiners. Although the losses on $2 \times 2$ splitter/combiners are negligibly small, 50% of the input radiation is lost in the nodes that operate based on $1 \times 2$ splitter/combiners due to modal radiation. Moreover, the nodes on the edges of the reservoir have a smaller number of neighbors compared with the nodes in the middle; therefore, their contribution to the optical mode mixing dynamics within the reservoir is limited. To circumvent these issues, a 4-port architecture [227] and Y-junction multimode combiners [228] that avoid the $1 \times 2$ splitter/combiners and enhance optical mode mixing dynamics have been proposed. Spatially-distributed nodes of an integrated photonic reservoir can also be implemented using crystal cavities [229], MRRs [230,231], or SOAs [119]. However, experimental verification and characterization of photonic reservoirs taking advantage of these technologies have yet to be demonstrated.

Although spatially distributing the optical computing nodes is the most straightforward and intuitive way to realize a photonic reservoir, it also poses some fabrication challenges. Alternatively, one can use a single nonlinear photonic node combined with

a time-delayed feedback line as a reservoir minimizing the hardware complexity of the photonic reservoirs. The delay-line-based reservoir computing concept was first developed by Appeltant *et al.* and demonstrated using electronic circuits [232]. Unlike the spatially distributed reservoir systems, this concept relies on a single hardware node that is multiplexed in the time domain with the help of a delayed feedback loop; hence, the reservoir nodes in these systems are sometimes referred to as virtual nodes or virtual neurons. While the delay-line-based single-node computing offers various advantages mainly due to its minimalistic approach regarding the hardware and fabrication requirements, it also intrinsically restricts the node topology within the reservoir to a circular graph where each virtual node interacts with only two neighbors. Furthermore, the idea of delay-line-based computing essentially involves a trade-off between space and time. Assuming reservoirs with $N$ nodes, the delay-line-based system must be processing $N$ times faster than its spatially distributed counterpart to compute a given task at the same speed. Despite these limitations, on the other hand, the concept has attracted a significant amount of attention within the optics and photonics community, as it offers a way toward implementing optical machine learning systems with only a single computational primitive that can be better optimized and more accurately controlled compared with photonic neurons in multinode, spatially distributed reservoir computing systems.

The experimental realizations exploiting the delay-line-based computing toward photonic processors have mainly been designed around optoelectronic ring oscillators (OEROs) [233–239]. In such systems, the optical part typically consists of a laser source, a long fiber spool providing the time-delayed feedback, and a broadband Mach–Zender modulator producing the nonlinearity in the form of $\sin^2$ (see Fig. 5(a)). The fundamental architecture shown in Fig. 5(a) has been initially demonstrated by Larger *et. al.* [233] and Paquot *et al.* [234]; whereas the former focused on its

Figure 5




**Delay-line-based photonic reservoir computing schemes.** (a) Building blocks of a typical optoelectronic single-node delay-line based reservoir computing system. The state variable in the feedback loop is $\boldsymbol{x}[n]$, corresponding to the voltage at the radio frequency input of the electro-optic Mach–Zehnder modulator. (b) Building blocks of a typical all-optical single-node delay-line based reservoir computing system. The state variable in the feedback loop is the complex envelope of the electrical field at the output of the laser E[$n$]. PC, polarization controller; MZM, Mach-Zehnder modulator; DL, delay line; PD, photodiode; EC, electronic coupler; c, circulator.

applications in speech recognition and time-series prediction, the latter demonstrated nonlinear channel equalization and nonlinear autoregressive moving average operations. Since these earlier studies, numerous advances on delay-line-based reservoir computing systems have been reported, including time-interleaved reservoirs processing multiple tasks [240], OEROs oscillating in wavelengths generated by a tunable laser source [237], and multiloop [241] as well as recurrent [242] reservoir computing architectures, where the output is reinjected into the reservoir. Moreover, Soriano *et al.* have experimentally demonstrated that using a six-level mask signal, instead of a binary square wave, can increase the robustness of delay-line optoelectronic reservoir computing systems against the quantization noise at the input and output. Duport *et al.* have developed a different approach to circumvent the noise and bit rate limitations of OERO-based reservoir processors and they developed a fully analog optoelectronic computing system [235]. For the application of these systems toward conventional machine learning tasks, e.g., speech recognition, the work of Larger *et al.* represents a major advancement, where they managed to reach the processing rate of one million words per second with only 0.04% and 0.6% word error rate on TI46 and AURORA-2 datasets by using a differential phase shift key (DPSK) controlled Mach–Zehnder demodulator along with two integrated optical phase modulators to generate non-local, nonlinear phase-to-intensity conversion dynamics within the reservoir.

Although the OERO-based reservoirs take advantage of photons in information processing, their processing speed is compromised by the electronic components used in the feedback loop. This limitation has been addressed by developing all-optical time-multiplexed reservoir computing systems [243–252]. In all-optical configurations, the electronic components on the feedback arm of the OERO reservoirs (filter and amplifier) are replaced by SOAs [244,251], semiconductor lasers [245,248,249,253], VCSELs [252], diode-pumped erbium-doped microchip lasers [250], quantum cascade lasers (QCLs) [254] and even passive elements, e.g., coherently driven passive fiber cavities [246] and semiconductor saturable absorber mirrors, [247] offering low-loss, power-efficient solutions (see Fig. 5(b)). Compared with the processing rates of their optoelectronic counterparts in the range of megabytes per second, the all-optical single-node reservoirs can offer significantly faster operation, achieving information processing rates around several gigabytes per second [245]. In addition, the faster operation also enables the system to create more virtual nodes within the reservoir for a given time delay [216].

The spatiotemporal mode mixing dynamics of waveguides and fibers have also been investigated as a third alternative to realize photonic reservoir computing systems. Mesaritakis *et al.* [255] numerically demonstrated that a single polymer waveguide could serve as a computing medium where the excited transverse modes act as the computational nodes of a reservoir. The input in their system is a light beam modulated with the help of a SLM coupled into the waveguide. Different modulation patterns over the light modulator excite various transverse modes supported by the waveguide geometry. According to their forward model, the air–polymer interface creates a short fiber cavity. In addition, they assumed a longer external cavity created by the use of beam splitter surrounding the waveguide. The presence of these two cavities ensures that the polymer waveguide, when used as a reservoir, exhibits fading memory. The output layer was implemented digitally following the detection of the speckle pattern generated by the superposition of all the excited modes of the waveguide with the help of a focal-plane array and an imaging system. A similar approach was reported in the experimental demonstration presented in Ref. [256], except that, instead of a polymer waveguide, the authors used a multimode optical fiber as the computing medium. They investigated the performance of their system in classifying audio signals from publicly available Japanese vowel dataset and reported 81.5% test accuracy, for which

a linear classifier can only achieve 43.2%. Finally, Tegin *et al.* expanded the fiber-based reservoir computing architecture to process both the spatial and spectral modes of short (10 ps) optical pulses modulated with the help of an SLM by using graded-index multimode fiber (GRIN MMF) and a lens-based projection system [257]. They experimentally demonstrated the performance of their system on regression tasks such as age estimation based on face images, classifying audio digits and x ray lung images for COVID-19 diagnosis.

### 3.2. Free-Space Optics and Engineered Diffractive Materials for Statistical Inference and Computing

In the previous section, we have presented and discussed the advances in optical computing and neural network architectures that are implemented based on integrated photonics. Although PIC technology has shown substantial progress addressing key design aspects, e.g., the implementation of all-optical nonlinear activation functions [121], toward realizing deep learning accelerators and neural networks, there are still some challenging engineering problems ahead to compete with electronic computing. As an example of some of these challenges, Zhang *et al*. [205] implemented the weights of the input and output layers of a three-layer neural network in the electronic domain because the effective number of interconnects on their photonic chip is not large enough to accommodate the entire inference task. This example points to the potential difficulties that might arise in the future of PICs toward realizing much wider and/or deeper neural network architectures while maintaining relatively compact footprints.

In an alternative approach, optical neural networks and the related computing systems can be placed directly in the path of propagating light waves before they are collected by an optoelectronic sensor. It is already known that FTs and fractional FTs can be performed using simple optical components, e.g., thin lenses, together with free space sections; this capability enables the all-optical implementation of some mathematical operations such as convolution and matrix–vector products [60,61,82]. With the development of some advanced material engineering and design modalities involving, e.g., metamaterials, plasmonics, and dielectrics, the associated fabrication techniques, e.g., two-photon polymerization [4,5] and additive manufacturing, along with the wide availability of GPUs, it has now become more feasible to precisely shape the optical wave field with task-specific, custom optical modulation surfaces without any physical wavefront recording as in previous holographic approaches [64,65], paving the way for efficient, parallel, fast, and scalable all-optical information processing systems [39,258,259]. Hence, intense computations that include machine learning tasks can be realized within a compact form factor, spanning, e.g., a few tens of wavelengths in the longitudinal direction. In Section 3.2.1, we present some recent results on free-space-based all-optical computation and statistical inference platforms that utilize diffraction of light. In addition to these all-optical processors and machine learning architectures, diffractive optical neural networks can also serve as front-end computing engines for the jointly trained hybrid (optical–electrical) network systems reducing the computational burden of the back-end electronic networks, potentially enabling advanced machine vision applications with low-pixel count, low-cost optoelectronic sensor arrays. We present some of the emerging results on such hybrid networks in Section 3.2.2.

#### *3.2a. All-Optical Inference and Computing Using Free-Space Optics and Engineered Diffractive Media*

The Rayleigh–Sommerfeld diffraction formulation has been shown to yield exact solutions for the propagation of optical waves inside an isotropic and homogeneous

medium both for far- and near-field diffraction [60,260,261]. According to this formulation, diffraction of light inside an isotropic and homogeneous medium (free-space) can be formulated as a shift-invariant linear system with an impulse response of

$$h(x, y, z) = \frac{z}{r^2}\left(\frac{1}{2\pi r} + \frac{n}{j\lambda}\right)\exp\left(\frac{j2\pi nr}{\lambda}\right), \tag{7}$$

where $r = \sqrt{x^2 + y^2 + z^2}$, $n$ and $\lambda$ are the refractive index of the medium and the wavelength of the propagating light, respectively, and $j = \sqrt{-1}$. The impulse response depicted in Eq. (7) satisfies the Helmholtz equations and represents the diffracted optical field in an exact manner within the assumptions of the scalar wave theory. If the target optical system operates within the small-angle regime, this exact diffraction formulation can be simplified and the light diffraction phenomenon can be approximated based on the Huygens–Fresnel principle [60]. Specifically, when the term $r = \sqrt{x^2 + y^2 + z^2}$ is represented based on the binomial expansion of the square root, it can be written as

$$r = z\left(1 + \frac{x^2 + y^2}{2z^2} - \frac{(x^2 + y^2)^2}{8z^4} + \ldots\right).$$

Ignoring the terms beyond the first two inside the parentheses, under the assumption that $\frac{x^2+y^2}{z^2}$ is small enough, leads to the following approximate free-space impulse response:

$$h(x, y, z) = \frac{e^{\frac{j2\pi}{\lambda}z}}{j\lambda z}\exp\left(\frac{j\pi}{\lambda z}(x^2 + y^2)\right) \tag{8}$$

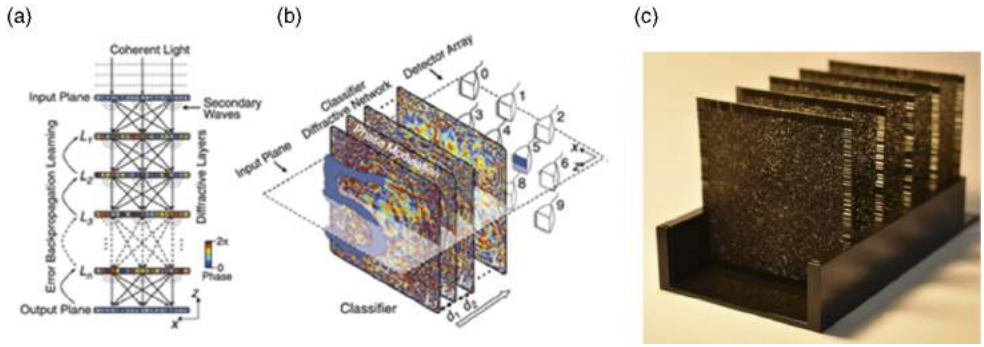
also known as Fresnel kernel [60]. Under this assumption, given a thin lens for which the optical modulation can be described as a quadratic phase function, a FT relationship between the two focal planes of the lens can be established. This Fourier transforming property of thin lenses, based on the Fresnel approximation, has been exploited to compute convolution operation and used as one of the building blocks of some of the earlier optical computing schemes leading to numerous applications, including road sign and face recognition systems [66,86,87,95]. On the other hand, this relation is only valid under Eq. (8) rather than Eq. (7). The difference is that whereas Eq. (7) computes the interaction between all the modes supported by free space, Eq. (8) assumes a limited spatial bandwidth under the Fresnel approximation. Limiting the space-bandwidth product of the propagating waves in an optical computing system would make it highly challenging to fabricate compact optical processors for large-scale, high-throughput computation over large fields of view.

A recently emerging optical computing platform without the use of traditional lenses is based on successive diffractive layers that are engineered and optimized to collectively compute a function through light-matter interaction. This framework is termed diffractive deep neural networks ($D^2NN$) [39,40,262–264] and it uses deep learning to design a series of passive diffractive surfaces to all-optically compute a given statistical inference task using a compact, thin optical platform that spans a few tens of wavelengths axially [39]. Given a machine learning task, $D^2NN$ formulates the problem from the perspective of devising a black-box optical processor that aims to approximate the desired input–output transformation function, e.g., classification of input objects (see Fig. 6). The forward model of this 3D black-box is described over the complex-valued transmittance coefficients, $t(x_i, y_i, z_i) = \alpha_i \exp(j\theta_i)$, of the diffractive features/neurons that occupy predetermined locations inside the computing volume, $(x_i, y_i, z_i)$, as shown in Fig. 6(a). These diffractive neurons are connected based on the light diffraction depicted by the impulse response in Eq. (7) (taking into account all

the propagating modes in space, without any paraxial approximation) and they are trained and optimized to compute a given inference task by collectively processing the light waves propagating through the diffractive optical network.

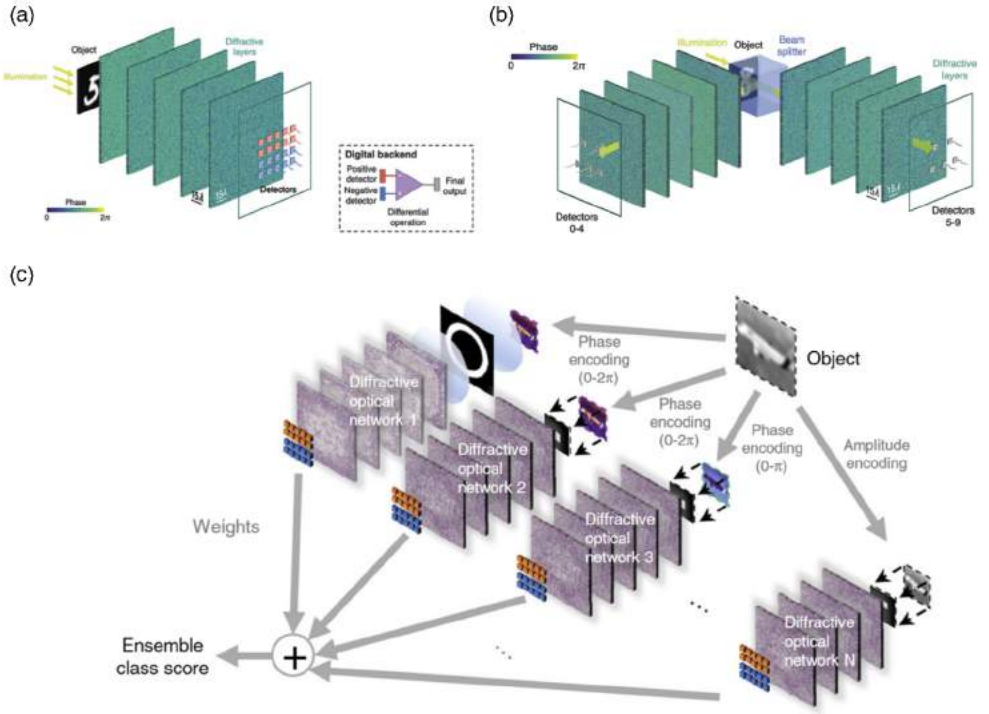
Similar to the training of electronic neural networks, at every iteration, a batch of inputs are virtually propagated through the diffractive optical network model using a computer environment, and the complex-valued transmittance of each diffractive neuron is updated according to the gradient of its parameters with respect to a loss function that is specifically tailored for the desired inference task. Once the training is finished using a computer, the transmissive (or reflective) surfaces constituting the diffractive network are fabricated to physically form the all-optical processor, which does not require any power to compute, except for the illumination light. The success of these physically formed, passive optical inference platforms was demonstrated using 3D-printed diffractive layers operating at terahertz wavelengths with, e.g., 0.2 million diffractive neurons fabricated over 5 successive layers [39,40,265] as illustrated in Fig. 6(c). These earlier demonstrations reported diffractive optical networks' generalization and statistical inference capabilities for object classification tasks. For example, blind testing accuracies as high as >98% and >90% have been reported to classify amplitude-encoded handwritten digits and phase-encoded fashion products, respectively [266]. To reach these classification accuracies for an optical computing hardware, Li *et al*. [266] implemented a series of design advances compared with the original $D^2NN$ architecture [39]. The most notable of these changes is a differential detection scheme shown in Fig. 6(a). The all-optical classification framework reported in Ref. [39] assigns one output detector to each data class in the target dataset, as shown in Fig. 6(b). The final inference decision is given based on the maximum signal detected among all the output detectors, i.e., max($\boldsymbol{I}$), where $\boldsymbol{I}$ represents the detected optical signals at the output plane of the diffractive network. With the differential scheme depicted in Fig. 7(a), Ref. [266] doubles the number of optical detectors at the output plane and assigns a pair of detectors to each data class representing the positive,

Figure 6




**Diffractive deep neural networks ($D^2NN$).** (a) The forward optical model of diffractive networks parameterize a given machine learning task as a function of the complex-valued transmittances of diffractive neurons on a series of layers connected to each other via light diffraction [39]. The amplitude and/or phase values of the transmittance over each neuron are trained using deep learning based on a task-specific loss function. (b) A diffractive all-optical handwritten digit classifier infers the data class of an input object (in this case the digit "5") by routing most of the incoming photons onto the corresponding class detector. (c) 3D-printed diffractive optical network that is fabricated for the experimental demonstration of all-optical object classification based on the $D^2NN$ framework. From Lin *et al*., Science **361**, 1004 (2018) [39]. Reprinted with permission from AAAS.

Figure 7




**Advances in the inference and generalization capacity of diffractive optical networks.** (a) Schematic of diffractive optical networks using differential detection scheme that assigns a pair of detectors to each data class. The class scores are computed based on the normalized difference of the optical signals collected by each pair of detectors [266]. (b) Schematic diagram of the class-specific diffractive optical network system design proposed in Ref. [266]. The classes of a target dataset are divided into subsets and assigned to different diffractive optical networks that are trained jointly. Each diffractive network in the system is trained to solve a simpler classification problem with reduced number of classes [266]. (a) and (b) Reprinted from Li *et al.*, Adv. Photonics **1**, 046001 (2019) [266]. Copyright 2019 SPIE. (c) Schematic diagram of an ensemble of diffractive networks, as proposed in Ref. [267]. The final ensemble class score is computed through a weighted summation of the differential detector signals synthesized by the individual diffractive networks within an optimized ensemble [267]. Reprinted by permission from Macmillan Publishers Ltd: Rahman *et al.*, Light. Sci. Applicat. **10**, 14 (2021) [267]. Copyright 2021.

$\boldsymbol{I}_+$, and negative, $\boldsymbol{I}_-$, parts of the final class scores. Accordingly, in this differential design the max operation is computed over the normalized differential optical signals, i.e., $\frac{\boldsymbol{I}_+ - \boldsymbol{I}_-}{\boldsymbol{I}_+ + \boldsymbol{I}_-}$. This simple adaptation in the output plane configuration brings significant improvement in the all-optical classification accuracies at the expense of using twice as many detectors at the output plane of a diffractive network; for example, for MNIST dataset, 20 detectors (10 $\boldsymbol{I}_+$ detectors and 10 $\boldsymbol{I}_-$ detectors) are needed for a differential $D^2NN$ design [266].

To further improve the inference capacity of diffractive optical networks, several works directly adapted a few related ideas and concepts from the machine learning literature, e.g., class-specific network training [266] (Fig. 7(b)), ensemble learning [267] (Fig. 7(c)), and skip network connections [268]. In the class-specific design scheme [266], the data classes in a target dataset are divided into subgroups. If, for example, the target dataset is handwritten digits (MNIST) with a total of 10 classes,

then these are separated into $P$ subgroups. Each subgroup is assigned to a different diffractive network, i.e., each diffractive network tries to solve a reduced classification problem with $10/P$ individual data classes. The class scores detected at the output plane of each diffractive network are combined to determine the network inference based on $\max(\boldsymbol{I})$ (see Fig. 7(b)). Compared with a single diffractive network-based inference, this class-specific design strategy with $P$=10 results in ~1.5%, 2.5%, and ~6% increase in the blind testing accuracy for the classification of digits, fashion products, and CIFAR-10 datasets, respectively.

In another work, Rahman *et al*. [267] applied ensemble learning techniques for designing a diffractive optical network system. Towards this end, 1252 different $D^2NN$ classifiers were trained for the more challenging task of classifying CIFAR-10 images. Unlike the previous class-specific design approach [266], these $D^2NN$ models were trained *independently*, i.e., without any feedback between any of the networks during their training. The diversity of the diffractive optical networks in the ensemble was accomplished by engineering unique spatial filters for each model. Each filter was placed between the input plane and the first trainable layer of the corresponding diffractive network (see Fig. 7(c)). Following the training phase, they performed a pruning algorithm to reduce the number of diffractive optical networks in the final system extracting optimized combinations of diffractive networks that can achieve improved classification accuracies with a limited number of models working together. Based on this ensemble learning scheme, blind testing accuracies of 61.14%±0.23% and 62.13 ± 0.05% were achieved with 14 and 30 diffractive networks selected through an evolutionary pruning algorithm [267], reporting the highest classification accuracy numbers for the CIFAR-10 dataset reported with a system based on passive diffractive networks.

Beyond diffractive optical networks that classify objects using light diffraction through passive materials, the $D^2NN$ framework has also found a plethora of other applications, including the categorization of human action images [269], multiview 3D object recognition [270], image segmentation, salient object detection [271], overlapping phase-object classification and image reconstruction [272]. Numerous research groups have shown successful experimental demonstrations of diffractive networks fabricated with different methods, operating at various parts of the EM spectrum. Recently, Goi *et al*. reported an application of the $D^2NN$ framework for the deep-learning-based design of an optical on-chip image encryption engine directly integrated on top of an infrared (IR) CMOS sensor [273]. They managed to create a neuron density of 500 million/cm$^2$ by fabricating diffractive neurons of size ~416 nm based on a nano-printing method that uses femtosecond lasers and two-photon polymerization [273]. Optical logic gates, such as NAND, AND, OR, and NOT have also been implemented all-optically using diffractive optical networks [274,275]. Reporting promising analysis on the cascadability of the $D^2NN$-based logic gates, Ref. [275] presented the design of an all-optical half-adder composed of five cascaded diffractive NAND networks, each with an identical design.

SLMs have also been used, instead of passive diffractive surfaces, within the $D^2NN$ framework leading to reconfigurable and adaptive designs. The main advantage of replacing the passive layers with dynamic electro-optic modulators is the adaptability to erroneous physical conditions through transfer learning [269] or *in-situ* backpropagation [276]. As a compromise, though, these dynamic electro-optic modulators increase the complexity and power consumption of the underlying diffractive network system compared to other solutions based on entirely passive components. There are also reported $D^2NN$ designs which use meta-atoms/metamaterials, e.g., $TiO_2$ nanopost on glass [277], as the computational precursor of a diffractive optical network;

however, an experimental diffractive network system based on metasurfaces (with sub-wavelength structures) is yet to be demonstrated.

The computational capabilities and inference/generalization capacity of diffractive optical networks composed of linear light modulation surfaces have been investigated in Ref. [263]. This theoretical analysis showed that for a diffractive optical network connecting an input field of view of size $N_i$ to an output field-of-view of size $N_o$, the dimensionality of the solution space of diffractive optical networks increases linearly, proportional to the number of diffractive neurons, $N$, working collectively within the diffractive system, up to a limit dictated by $N_iN_o$. The same analysis also showed that if all $N$ neurons reside over a single diffractive layer, the efficiency of deep-learning-based training decrease substantially compared with the case where $N$ neurons are distributed over two or more diffractive surfaces that are successively placed. In other words, for a given statistical inference task, e.g., classification of CIFAR-10 images, diffractive optical networks with a larger number of layers can achieve higher blind testing accuracy, diffraction efficiency, and optical signal contrast [263], despite using linear optical materials. This behavior is also intuitively similar to the case in electronic neural networks. Although the universal approximation theorem proves that a neural network having a single hidden layer (with an appropriate nonlinear activation function) is a universal function approximator, in practice the gradient-based learning has a hard time finding the optimal parameters for shallow networks. In addition, the number of neurons required in a shallow network architecture increases exponentially to reach the same generalization performance of a deeper network. Therefore, in both electronic networks and diffractive optical networks, given a certain number of neurons, it is a better design practice to distribute these available neurons over deeper network architectures [263,278].

In addition to all-optical statistical inference tasks, the $D^2NN$ framework can also be utilized to design diffractive optical networks performing various deterministic all-optical computing tasks [278–280]. For applications in optical communications, Huang *et al*. [280] designed diffractive optical networks that can achieve three modes of optical modulation, namely, orbital angular momentum-shift keying (OAM-SK), OAM multiplexing and demultiplexing, and OAM-mode switching. Kulce *et al*. [278], on the other hand, reported $D^2NN$-based linear transformations and demonstrated that the deep-learning-based training of diffractive optical neural networks can be used to synthesize an arbitrary, complex-valued linear transformation between an input and an output field of view. Some of these arbitrarily selected linear transformations performed using diffractive networks included unitary, non-unitary, and non-invertible randomly generated complex-valued matrices, 2D discrete FT, 2D permutation operations, and high-pass filtered imaging, highlighting the broad scope of all-optical computing applications that can benefit from diffractive networks.

Beyond enabling coherent all-optical processors that utilize both the phase and amplitude information of light, the $D^2NN$ framework has also been extended to process temporally incoherent, broadband light to compute machine learning tasks all-optically [40]. The broadband computational capabilities of diffractive optical networks might potentially bring deep-learning-driven solutions addressing problems pursued for a long time in the field of computational imaging and sensing fields. One example is single-pixel machine vision. Although there are various approaches toward designing single-pixel imaging systems, these often rely on time-multiplexed data collection schemes, mechanical and/or electrical scanning, and computationally expensive information recovery algorithms, hindering their practicality and utilization in many applications. Incorporating the material dispersion properties into the deep-learning-based forward training model, Li *et al*. [40] have demonstrated diffractive optical networks that enable single-shot, single-pixel machine vision systems by

simultaneously processing multiple wavelengths to classify objects, e.g., handwritten digits [40]. In this single-pixel $D^2NN$-based machine vision framework, the broadband input light transmitted through (or reflected by) an input object is processed by specially designed diffractive layers trained to encode the spatial information of input objects into the intensity of the spectral components collected by a single-pixel detector aperture located at the output plane, where each wavelength represents one data class. The success of this framework has been experimentally demonstrated by classifying handwritten digits all-optically with a single pixel. Blind testing inference accuracies as high as 96.82% were reported with this $D^2NN$-enabled spectrally encoded single-pixel machine vision system. The broadband light processing capabilities and the deep-learning-based, data-driven training of diffractive optical networks have also been used for solving inverse optical design problems such as spatially-controlled wavelength-demultiplexing [41] and optical terahertz pulse shaping [42], which is discussed in more detail in Section 4.1.5.

One of the key areas that would significantly enhance the inference capacity and function approximation power of diffractive optical networks is potential nonlinear activation functions that can be embedded in the signal pathway between the input and output fields of view. For this aim, there have been some emerging approaches to introduce nonlinearity in diffractive optical networks [268,269,271,281]. Some of the proposed implementations of optical nonlinearities in Refs. [268,269,271] rely on photodetectors and/or sensor arrays. They also require the use of dynamic, reconfigurable electro-optic modulation devices, which might result in an inferior computational speed. Zuo *et al*. [281], on the other hand, has shown an all-optical nonlinear activation function based on electromagnetically induced transparency (EIT), which is a light-induced quantum interference effect among atomic transitions. This nonlinear neuron activation platform is based on $^{85}Rb$ atoms in a 2D magneto-optical trap. These atoms were maintained in the ground state, $|1$ and the atom cloud was illuminated with two laser beams propagating in opposite directions. Although the coupling laser beam, which represents the input to the neuron, is set to be in resonance with the atomic transitions $|2 \rightarrow |3$, the counterpropagating probe beam is in resonance with $|1 \rightarrow |3$ transitions. In the absence of a coupling beam, the atom cloud is opaque for the transmission of the probe beam. When the coupling beam is present, the relation between the intensity of the coupling beam, $I_{in}^{c}$, and the output probe, $I_{out}^{p}$ can be written as

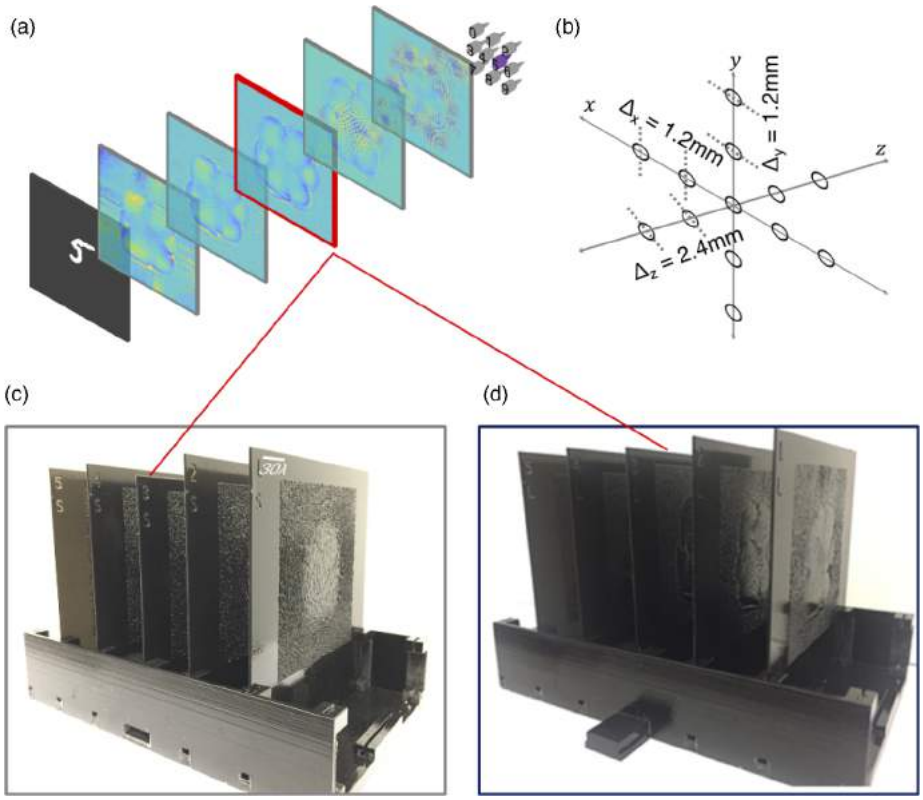
$$I_{out}^{p} = I_{in}^{c} \exp\left(-OD \frac{4\gamma_{12}\gamma_{13}}{\Omega_c^2 + 4\gamma_{12}\gamma_{13}}\right), \tag{9}$$

where $OD$ and $\gamma_{ij}$ denote the atomic path depth for the transition $|1\rangle \rightarrow |3\rangle$ and dephasing rate between the states $|i\rangle$ and $|j\rangle$, respectively. The term $\Omega_c$ represents the coupling field Rabi frequency, and its square is proportional to the coupling laser intensity. The development of such all-optical nonlinear activation models that are much easier to fabricate and align/maintain, scalable, and efficient, is of particular importance for the future of optical computing platforms but remains an important engineering challenge in the field.

Physical implementation errors and sources of noise present yet another major challenge in these all-optical computing systems and diffractive neural networks. Although this is a common problem affecting all analog computing devices including, e.g., electronic VLSI systems, there are some error sources specific to the optical computing systems using wave propagation and engineered media, such as mechanical misalignments and fabrication inaccuracies/imperfections. One way to tackle these potential issues that would prevent the diffractive optical networks and other free-space optical computing systems from reflecting their true performance in practical applications is

to replace the passive modulation surfaces with reconfigurable electro-optic modulators that can be calibrated *in situ* [269,276]. On the other hand, the use of dynamic modulators could result in a significant increase in the system complexity, cost, and power consumption. Alternatively, it has been shown that the evolution of diffractive surfaces during the deep-learning-based training of a diffractive network can be regularized and guided toward diffractive solutions that can maintain the inference accuracy despite mechanical misalignments [265]. This misalignment-tolerant diffractive network training strategy models the layer-to-layer misalignments, e.g., translations in $x, y, z$, over random variables and introduces these errors as part of the forward optical model, inducing "vaccination" against such system inaccuracies and/or variations. In their proof-of-concept experiments, the authors fabricated and compared a vaccinated diffractive network with a non-vaccinated one, trained for the classification of handwritten digits. As shown Fig. 8, misaligning the 3rd diffractive layer in a 5-layer

Figure 8




**Experimental demonstration of vaccinated diffractive optical networks (v-D$^2$NNs) [265].** (a) Schematic diagram of a diffractive optical network that is vaccinated against *both* lateral and axial layer-to-layer misalignments. (b) The positions of the center of the 3rd diffractive layer during the experimental testing. The central location corresponds to ideal placement (zero misalignment), whereas the remaining 12 positions represent various degrees of misalignment of the 3rd layer with respect to the others. (c) 3D-printed unvaccinated design that failed to infer the correct data class 23 times out of in total 78 measurements over 6 different test objects (13 positions of the 3rd layer for each object). The 3D-printed, vaccinated design shown in (a) that failed only twice for inferring the input object class, meaning that in the remaining 70 physically misaligned network constructions, the final class assignments were all correct, demonstrating the success of "vaccination." Reprinted from Mengu *et al.*, Nanophotonics **9**, 4207 (2020) [265]. Copyright 2020 De Gruyter.

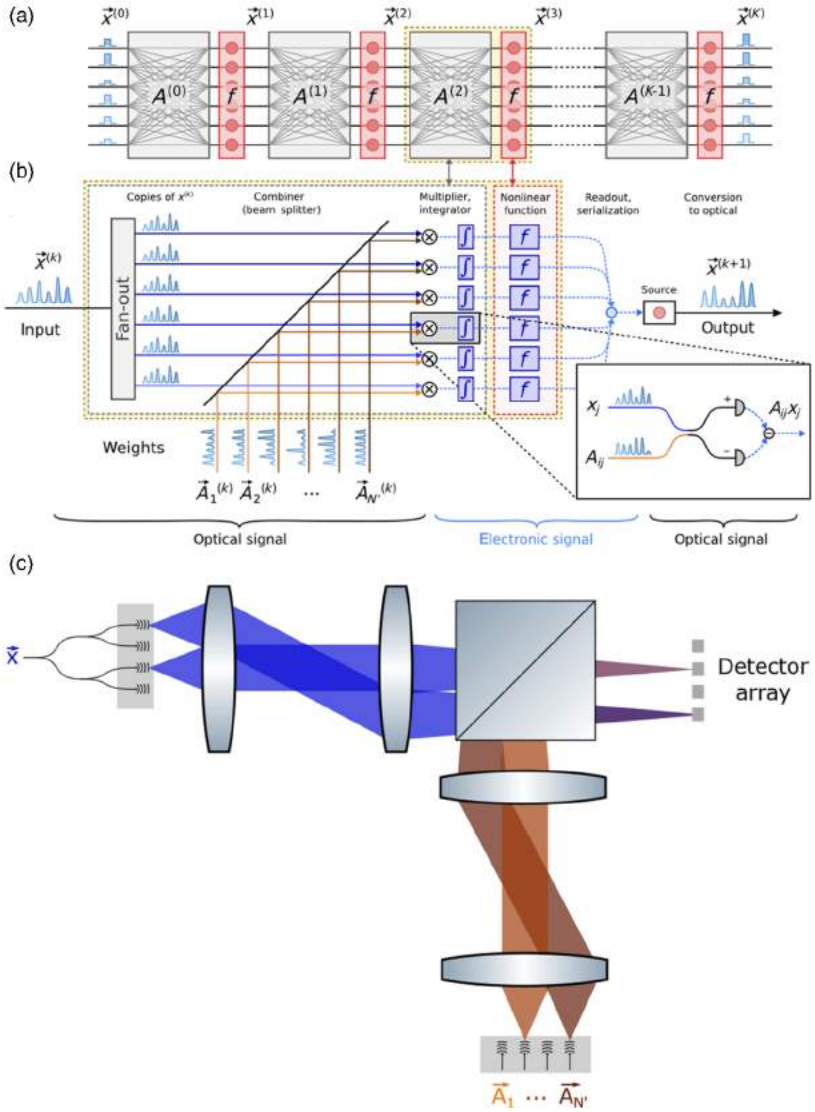
all-optical classification network to 12 different locations around its ideal position, the authors measured the intensities collected by the class detectors at the output plane of the diffractive network for 6 different input test objects/digits never seen by both of the networks. Although both of these diffractive network designs predicted the object classes correctly when the 3rd diffractive layer is at its ideal location, in the remaining 72 measurements, the non-vaccinated diffractive network failed to reveal the correct class 23 times, whereas the vaccinated network managed to infer the correct object class in 70 measurements, failing only twice, highlighting the efficacy of the vaccination strategy. The same training scheme can also be extended to mitigate the effects of other potential error sources, e.g., fabrication inaccuracies [282] and optoelectronic detection noise, improving the robustness of diffractive networks toward practical applications without requiring bulky and expensive electro-optic modulator systems. Based on a similar training approach, diffractive optical networks composed of passive layers can also adapt to random variations at their input in the form of object scaling, shift, and rotation [283].

$D^2NN$ framework has introduced a unifying perspective on deep learning, wave optics, and light–matter interactions and, with its broad applicability, it has fueled a large number of research efforts as outlined previously. On the other hand, in addition to diffractive optical networks, there exist a variety of recently proposed optical computing techniques that also exploit waves within engineered media. For instance, the wave physics of an inhomogeneous medium can act as an analog RNN [284]. By optimizing the spatial distribution of two or more types of materials based on an adjoint method [35], it is possible to use wave dynamics to classify vowels mimicking a RNN [284]. A similar approach has also been shown in [285], where the authors designed a nonlinear nanophotonic medium of size 80λ by 20λ. By numerically solving Maxwell's equations and optimizing the locations of air holes within the medium, the authors demonstrated a system achieving ~79% blind testing accuracy for the classification of handwritten digits.

In a recent work, Hamerly *et al*. proposed another innovative approach to optical computing. They implemented ultra-low-power matrix–vector multiplication architecture using two lens-based imaging systems combined over a beam splitter and balanced homodyne detection [286]; see Fig. 9. In their experimental system, the input vector $\boldsymbol{x}$ and the weight matrix $\boldsymbol{A}$ are generated by a master laser feeding two grating antennas, i.e., point sources, on two arms of a beam splitter. The entries of the input vector $\boldsymbol{x}$ are encoded in the complex-valued amplitudes of a pulse sequence, and it is replicated over an array of point sources (see Fig. 9(b)). On the second arm of the beam splitter, each point source is used to create the rows of the matrix also encoded as a time-domain pulse sequence. The light waves coming from these point sources are superposed over a beam splitter, as shown in Fig. 9(c). For instance, in the case of a vector–vector inner product with one point source on the arm corresponding to the weight matrix $\boldsymbol{A}$, for a given time point $j$, the field on the beam splitter can be written as $x_j + A_{ij}$. This field is then projected on two detectors creating the intensities, $I_+ = 0.5\|x_j + A_{ij}\|^2$ and $I_- = 0.5\|x_j - A_{ij}\|^2$. The difference between these intensity values can be written as $I_+ - I_- = 2Re[A_{ij}^* x_j]$, which is proportional to the multiplication of the weight matrix and the vector entries. An important design consideration in such an optical computing scheme is that the total path length from the grating antennas to the detectors must be shorter than both the coherence length of the laser and the $c\tau_{pulse}$, where $c$ and $\tau_{pulse}$ denote the speed of light and the duration of a pulse in the time sequences representing the entries of $\boldsymbol{x}$ and rows of $\boldsymbol{A}$.

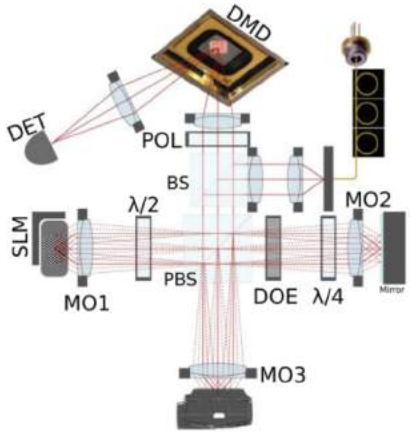
Beyond these neural network implementations, diffractive surfaces and free-space wave propagation have also been harnessed to realize optical reservoir computing

Figure 9




**Implementation of a large-scale optical neural network based on imaging optics and homodyne detection** [286]**.** (a) Feedforward neural network architecture with $K$ layers. (b) The proposed implementation of the multiply-accumulate (MAC) operation (matrix–vector multiplications) on each layer based on time-multiplexing and replicating the entries of the input vector $\boldsymbol{x}$ over a series of point sources driven by the same master laser. Each row of the matrix $\boldsymbol{A}$ is converted to a time-domain signal emitted by a nanoantenna/point source, therefore, the number of point sources on the matrix arm of the beam splitter is equal to the number of rows. The matrix–vector multiplication is achieved through homodyne detection at the output plane. According to the proposed optical network design scheme, nonlinear activation is applied electronically following the homodyne detection. (c) Two imaging systems are combined over a beam splitter to superpose the optical signals representing the input vector $\boldsymbol{x}$ and the weight matrix $\boldsymbol{A}$. For a given time point, homodyne detection at the output detector array generates signals that are proportional to the multiplication of the entries in vector $\boldsymbol{x}$ and the weight matrix $\boldsymbol{A}$ **,** i.e., $A_{ij}^{*}x_j$; consequently, time integral of the collected signal at the $i$th homodyne detector reveals the inner product between the $i$th row of matrix $\boldsymbol{A}$ and the input vector $\boldsymbol{x}$. ). Reprinted under a Creative Commons license.

Figure 10




**Implementation of a reservoir computing system based on the diffractive coupling of wave fields in free space** [225]**.** The state of the reservoir encoded on the SLM is imaged onto the camera (CAM) through a polarizing beam splitter (PBS) and a diffractive optical element (DOE). A digital micromirror device (DMD) creates a spatially modulated image of the SLM's state corresponding to the output layer connections in a reservoir system i.e., $\boldsymbol{W_{DMD}} = \boldsymbol{W_{out}}$. The field right after the DMD is focused onto a detector integrating the weighted internal reservoir state to generate the output signal [225]. Reprinted with permission from [225]. Copyright 2018 Optical Society of America.

systems. Compared with their counterparts implemented through PICs, time-delay feedback lines, and waveguides, the generation of a reservoir through the diffractive spatial coupling of propagating wave modes provides a significantly higher number of computational neurons to be accommodated. Although most delay-line-based systems typically operate using a few hundred neurons, diffractive systems can provide more than 10,000 nodes [287]. On the other hand, most of the implementations reported to this date have relied on electronic feedback mechanisms and digitally implemented nonlinear activation functions [287–289], significantly hindering the computational speed and power consumption of these systems. To offer a solution to this issue, Ref. [224] investigated the coupling dynamics of a reservoir that has 64 nodes in the form of an $8 \times 8$ VCSEL array, a DOE, and a reflective SLM. They set the separation between the DOE and the SLM to ensure spatial overlap between the higher diffraction orders and the main order generated by the DOE over the surface of the light modulator. The use of a reflective light modulator creates the optical feedback mechanism required by the reservoir computing, and in one loop, the light emitted by the VCSEL array impinges over these sources after making a double pass through the DOE. While utilizing one-half of the SLM for controlling the reservoir states, the other half was used for the weighted integration of output signals focused onto a detector.

Recently, the concept of creating internal reservoir connections based on diffractively coupled light sources has been extended to accommodate ~2000 neurons for its application in reinforcement learning [225] (see Fig. 10). Unlike the previous approach, the authors did not use a VCSEL array, instead fed the information into the system through a single node. According to their system design shown in Fig. 10, the internal reservoir connections was realized by the light beam making a double pass through a DOE, hence, $\boldsymbol{W_{res}} = \boldsymbol{W_{DOE}}$. The state of the reservoir, on the other hand, was defined as the optical intensity over the polarizing beam splitter (PBS). The output layer connections, $\boldsymbol{W_{out}}$, were implemented using a digital micro-mirror device (DMD).

They demonstrated the success of their system by performing Mickey–Glass time-series prediction based on 2025 reservoir nodes. Although they managed to achieve impressive NMSE values for the application of time-series prediction, direct use of the optical reinforcement learning architecture shown in Fig. 10 in practical applications could be challenging due to its power consumption, cost, complexity and ~5 Hz system update speed.

Although these all-optical information processing and neural network architectures that take advantage of free-space optics generally allow low-latency, low-power, and highly parallel computing capabilities, their current inference performance is largely crippled by the lack of practical, low-power, and scalable nonlinear activation functions as discussed earlier and emphasized in the literature (see, for example, the Supplementary Materials of [39]). Therefore, their generalization capabilities might be insufficient for the end-to-end all-optical computing of more complex machine learning tasks, e.g., multitask inference systems. Some of these more challenging cases can be addressed by integrating these optical computing systems as analog front-end processors with electronic (back-end) neural networks, creating hybrid (optical–electronic) systems working in collaboration. On the other hand, the formation of hybrid systems can also benefit electronic neural networks that have already become an integral part of modern-day machine vision systems. Specifically, with task-specific, trainable diffractive optical surfaces replacing the standard lens-based imaging optics used in cameras, next-generation machine vision systems could become more resource-efficient and faster. This unique opportunity is discussed in the next subsection.

#### *3.2b. Optical Neural Networks as Analog Front-End Processors Integrated with Electronic Back-End Neural Networks for Hybrid Machine Vision Systems*

In a conventional machine vision system, there are three fundamental building blocks: (1) a lens-based imaging optics (front-end), (2) an optoelectronic sensor array, and (3) electronic neural networks (back-end). Among these different parts, there is a clear division of labor in which the optical part directs the incoming light to form the image of a scene that is collected and digitized by the optoelectronic sensor array. Finally, the information is digitally processed for various inference tasks by subsequent electronic neural networks. From a machine learning perspective, the same components can also be interpreted as an encoder–decoder system [103]: the optical part synthesizes a representation of the scene, its image, and the electronic neural network extracts, processes and/or decodes the information to achieve the desired machine learning task(s). Based on this point of view, the intensity of the light field collected by a focal-plane array does not have to be the exact replica of what a human eye would see or interpret, but rather it should correspond to a representation of the input that can be decoded by a jointly-trained back-end electronic network.

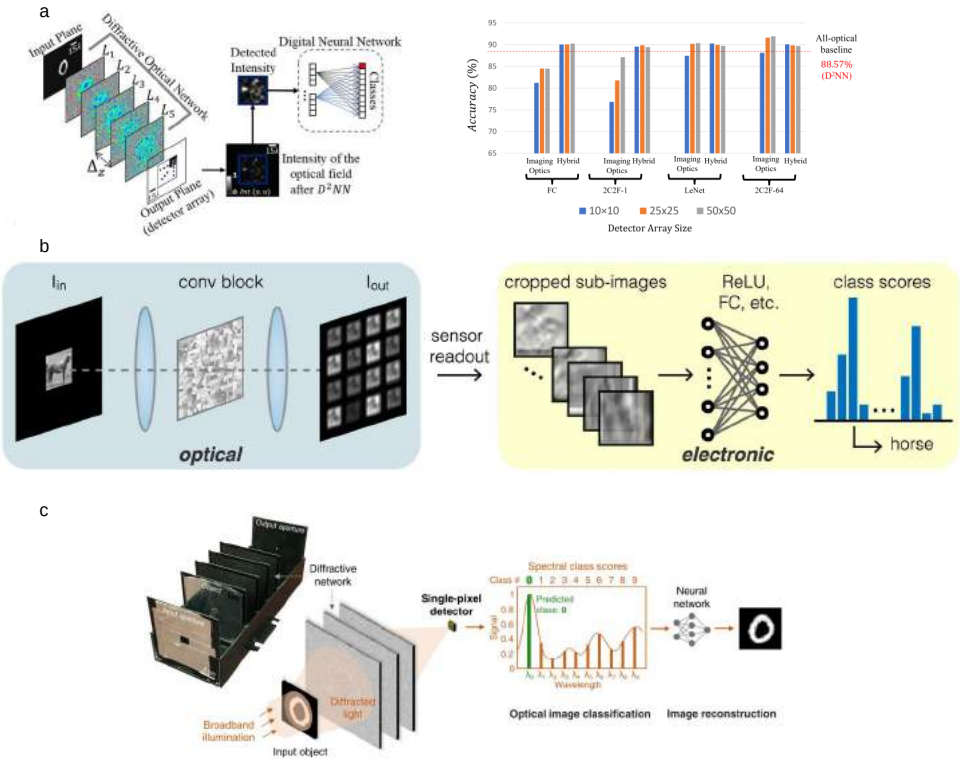
From this perspective, the optical computing systems replacing the traditional imaging optics can be seen as front-end processors that work alongside the back-end electronic neural networks for improving, for example, the computational speed, frame rate, memory requirements, and power consumption of machine vision systems. Over the past few years, several hybrid network systems have been presented demonstrating object recognition and/or task-specific computational imaging tasks. Examples of object classification systems, where the optical front-end network significantly reduces the resolution requirement on the focal-plane array and the computational burden on the back-end electronic network is presented in Ref. [262]. By jointly training a diffractive optical network with an electronic neural network, the authors have shown that it is possible to reduce the number of pixels at the focal plane array by approximately a factor of eight, without any compromise on the inference accuracy compared with a conventional vision system with an ideal, aberration-free imaging

optics (see Fig. 9(a)). According to the reported results, for instance, LeNet [59] can classify the images of fashion products with an accuracy of 90.33% based on an ideal imaging optics and an optoelectronic sensor having the full resolution of $28 \times 28$ pixels. On the other hand, if the number of pixels on the focal-plane array is reduced to $10 \times 10$ (approximately a factor of eight reduction), the related loss of information due to undersampling and aliasing hurts the inference performance of LeNet, which can only provide 87.44% classification accuracy. When the same LeNet architecture is jointly trained with a 5-layer diffractive optical front-end, the inference accuracy with the same low-resolution sensor (10×10 pixels) remains to be >90.2%.

In addition to allowing the use of low-resolution sensors, a diffractive optical network front-end also contributes to the inference tasks and enables the use of shallower electronic neural networks at the back-end with reduced number of neurons and MAC operations to achieve the same inference accuracy. For instance, in the case of a $25 \times 25$-pixel focal-plane array, a hybrid system composed of a 5-layer diffractive network and a *single-layer* fully connected network can provide 90.08% inference accuracy for classifying fashion products [262]. On the other hand, a deeper network, LeNet, can achieve 90.19% classification accuracy by processing images created by an ideal imaging system and a focal-plane array with the same resolution ($25 \times 25$ pixels). When we compare the computational cost of these two electronic network models, the single-layer shallow network only has $25 \times 10^3$ trainable parameters and computes the inference task with $50 \times 10^3$ floating-point unit arithmetic operations (FLOPs), whereas LeNet, has $60.8 \times 10^3$ learnable parameters computing the data class based on $\sim 10^6$ FLOPs. Therefore, when a diffraction-limited imaging system is replaced with a diffractive optical network that is jointly trained with a back-end, shallow electronic network, the memory usage, computational speed, and power consumption of the system can be reduced significantly by simplifying the back-end architecture [262].

In a hybrid machine vision system, the optical part can also be tasked to accompany the back-end electronic network as a co-processor, either assisting the electronic network by directly acting as its first layer [290] or completing one of the tasks all-optically in a multitask machine learning system [40,272]. Figure 9(b) illustrates an optical 4*f* correlator architecture proposed by Chang *et al*. to implement a given convolutional layer of an electronic neural network in the optical domain by processing spatially incoherent light [290]. In their hybrid optoelectronic system, they trained a neural network model, using a computer, composed of a convolutional layer with 8 filters of size $9 \times 9$ followed by a ReLU nonlinearity and a fully connected layer digitally mapping the output feature space of the convolutional layer to class scores. Owing to the non-negativity constraint of optical intensity, in their forward training model, they represented the 8 feature channels at the output of the convolutional layer as the difference between 16 channels with 8 positive and 8 negative channel pairs, termed as pseudonegative convolution. At the end of the training, 16 stacked filters of this pseudonegative convolutional layer were tiled side-by-side over a $4 \times 4$ grid providing the desired point spread function (PSF) of the 4*f* correlator as shown in Fig. 11(b). In the second step of their design, they optimized a phase mask that can achieve the desired PSF mimicking the target convolution operations. For the optimization of the phase mask, they also utilized a gradient-descent-based algorithm during which the height profile of the mask material is iteratively updated. As, at the end of convergence of phase encoding, the predicted PSF of the phase mask might not exactly match the ground truth, they applied additional training on the subsequent fully connected electronic layer to compensate for this discrepancy. Based on the outlined design procedure, their final forward model, including the phase-encoded PSF, predicted 51.0% accuracy for the classification of CIFAR-10 images. The reported experimental

Figure 11




**Hybrid (optical–electronic) neural network implementations.** (a) Jointly trained diffractive optical front end and electronic neural networks (back end) for the task of object classification [262]. Left, the concept of jointly trained diffractive and electronic neural networks toward achieving a machine learning task. Right, the performance of a co-trained hybrid system is compared against the classification performance of a conventional vision system which uses lens-based, diffraction-limited imaging optics with the same electronic neural network architecture. © 2019 IEEE. Reprinted, with permission, from Mengu *et al.*, IEEE J. Select. Topics Quantum Electron. **26**, 1–14 (2019) [262]. (b) Hybrid two-layer neural network where the first convolutional layer (9×9, 8 filters) is computed all-optically followed by the second fully connected layer in the electronic domain [290]. Reprinted by permission from Macmillan Publishers Ltd: Chang *et al.*, Sci. Rep. **8**, 12324 (2018) [290]. Copyright 2018. (c) A single-pixel multitask machine vision framework based on hybrid neural network systems. The diffractive optical front end encodes the spatial information of an input object into the intensity of predetermined spectral components collected by a single detector at the output plane. These spectral intensity values represent the class scores and the all-optical class inference is revealed by the *maximum* of these spectral intensity values. A shallow electronic back-end network is trained *separately* based on these optically synthesized spectral intensity values to reconstruct/recover the unknown image of the input object. The resulting multitask, single-pixel hybrid machine vision system all-optically classifies objects by utilizing the broadband processing and statistical inference capabilities of diffractive networks and also electronically reconstructs the images of the objects within the input field of view [40]. From Li *et al.*, Sci. Adv. **7**, eabd7690 (2021) [40]. Reprinted with permission from AAAS.

accuracy, on the other hand, was slightly lower (44.4%) due to the unwanted physical error sources causing the system conditions to deviate from the constructed forward model, highlighting the importance of noise mitigation strategies in optical computing systems [265,269,276,282].

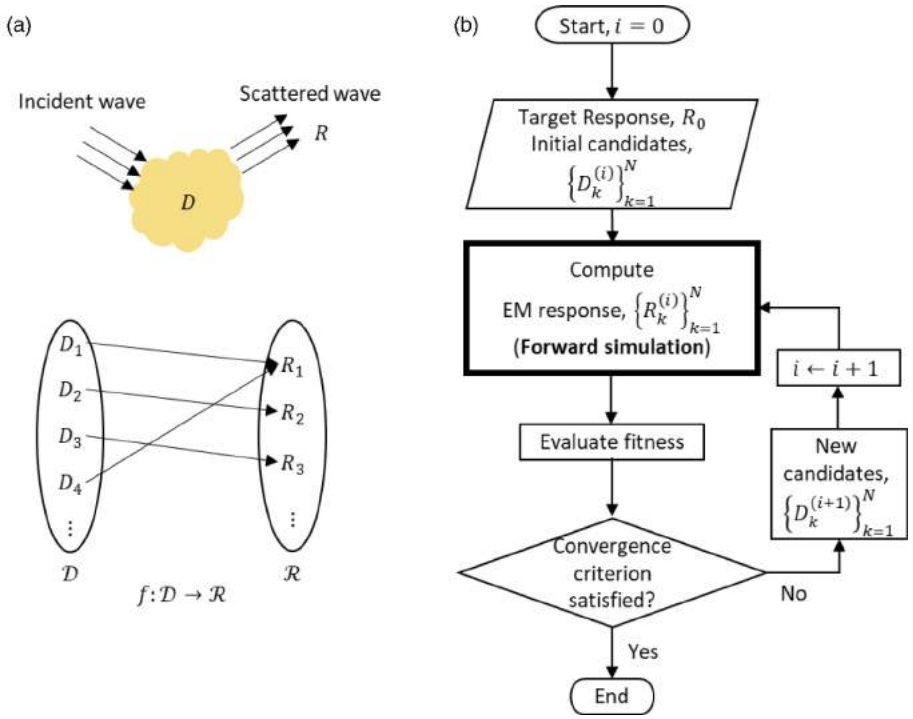
Figure 9(c) depicts another example of a multitask hybrid machine learning system where the optical front-end and the electronic neural network are trained for different tasks separately. In this particular case presented in Ref. [40], the diffractive optical network extracts the information on spatial features of an input object and encodes it into the intensities of a set of predetermined wavelength components that are collected by a plasmonic nano-antenna-based single-pixel detector [291]. The intensity values of the detected spectral components at the predetermined wavelengths represent the class scores. Thus, the optical front-end completes the task of image classification all-optically based on the spectral power distribution at a single pixel. Upon convergence of the training of the optical part, a back-end electronic neural network is trained *separately* to decode the highly compressed spatial information at the single-pixel output. Stated differently, this image reconstruction electronic network performs a task-specific data/image decompression, with the task being the recovery of handwritten digit images, establishing a multitask computational vision system that covers both electronic image recovery and all-optical image classification through a single-pixel machine vision system. A similar multitask, separately trained hybrid neural network system has also been presented in Ref. [272], where the authors took advantage of the coherent processing capabilities of diffractive optical networks to all-optically classify two spatially overlapping phase objects despite strong phase ambiguity due to spatial overlap at the object plane, solving an unconventional machine learning problem. They further demonstrated that back-end electronic networks trained based on the intensities collected by the class-specific detectors could, in fact, reconstruct the phase images of both objects addressing an inverse problem with, in general, a non-unique solution space [272].

## 4. DEEP LEARNING FOR DESIGN IN OPTICS AND PHOTONICS

### 4.1. Deep-Learning-Enabled Inverse Design for Optical and Photonic Devices

Deep learning has also been used extensively for the design of conventional optical components, e.g., inverse design of lens groups based on geometrical optics. For example, Côté *et al.* used deep learning as a tool to infer lens design starting points directly from the desired specifications, such as focal length, $f$ number, and field of view [292]. They trained a multilayer fully connected network that can automatically produce high-quality starting points by extrapolating from known designs. However, this framework was only used on simple air-spaced and cemented doublets, limiting the design space of the lens group. In a follow-up work, the same research group overcame this limitation by using a RNN that can dynamically capture the sequential structure of lens designs with a flexible number of elements, which allows a single model to adopt a shared representation of various lens structures that differ by the sequence of glass elements and air gaps so that the resulting model can generalize to new lens design structures for which there is no reference lens design [293]. In a recent work, they further improved their framework to allow the aperture stop to be placed anywhere in the generated lens starting point [294]. However, as promising as deep learning can be for the design of traditional ray optics-based devices, the bulk of the literature related to the applications of deep learning for photonic design is concerned with the task-specific inverse design of nanophotonic structures [295]. Starting with the following subsection we focus on these emerging applications of deep learning for inverse design in nanophotonics.

Figure 12




**Forward and inverse problems in nanophotonics.** (a) Forward simulation to compute the electromagnetic response $R$ for a design $D$ can be thought of as evaluating a function $f$ that maps the designs in the space $\mathcal{D}$ to their corresponding responses in the space $\mathcal{R}$. For almost all problems of practical interest, this is done by numerically solving Maxwell's equations with an electromagnetic solver. (b) An algorithmic flow chart representing conventional iterative approaches to inverse design. The major step in an iteration of such inverse design is the forward simulation of candidate designs for computing their electromagnetic responses, which is highly resource intensive.

#### *4.1a. Conventional Inverse Design Approaches Used in Nanophotonics*

The response $R$ of an EM system to incident radiation depends on the system geometry $D$, e.g., the spatial distribution of permittivity and permeability throughout the system. Computing the response of a system with given geometry to incident radiation, i.e., evaluating the function $f$ that maps system designs in the space $\mathcal{D}$ to their EM responses in the space $\mathcal{R}$ (forward computation/simulation; Fig. 12) entails solving the Maxwell's EM equations. Various analytical and numerical methods have been demonstrated for providing the approximate solutions to Maxwell's equations. The level of computational resources required for numerically solving Maxwell's equations depends on the complexity of the nanophotonic geometry as well as the desired accuracy of the solution. As useful as the forward simulation is, for engineering applications, it is often necessary to reverse engineer a desired response, i.e., to accurately find out a geometry $D$ that would give rise to a desired response $R$. Researchers in the nanophotonics field are frequently faced with inverse design problems, i.e., determining the nanostructure geometry that would give rise to a desired optical response [35]. The solution to nanophotonic inverse problems has often been guided by the intuition of expert practitioners or metaheuristic optimization approaches.

One of the basic approaches to solving inverse problems is possibly the "grid search," which involves solving the forward problem for many design points in $\mathcal{D}$ and selecting

the design $D$ that results in the nearest approximation to the desired response in terms of a well-chosen fitness metric/objective function. More sophisticated approaches toward inverse design problems utilize various additional optimization algorithms. The vast majority of these numerical optimization algorithms applied to EM inverse design problems are metaheuristic in nature, such as evolutionary algorithms, particle swarm optimization, and ant colony optimization [296]. These algorithms are the preferred choice in cases where the gradient of the objective function with respect to design parameter space is difficult or impossible to compute. Gradient-based optimization methods such as gradient descent and conjugate gradient method can be utilized if the gradient computation is practically feasible, for example, by using adjoint simulation [297].

An algorithmic flowchart that represents many of these approaches is depicted in Fig. 12. These algorithmic approaches to inverse design have one thing in common: all of them require solving/executing the forward model numerous times during the optimization process. As a result, even with adequate computational resources, the time required for finding a decent solution/approximation can become prohibitively large. Moreover, the results of all the intermediated forward simulations are, in general, not reusable. Another limitation of these approaches is that they are feasible for finding solutions only within a family of designs parameterized by a few design parameters.

#### *4.1b. Deep-Learning-Based Methods for Inverse Design in Nanophotonics*

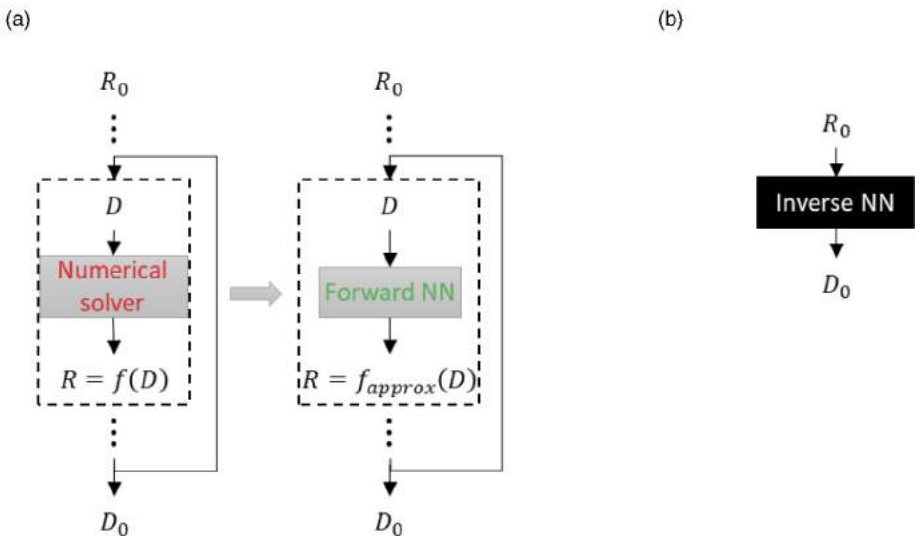
The ability of deep neural networks to approximate any function [50] has motivated nanophotonic researchers to exploit them in solving inverse design problems. The earliest use of neural networks in EM inverse design dates back to the 1990s for microwave applications [298,299]. A review of the works in deep-learning-based inverse design prior to the onset of the century can be found in Ref. [300]. Since 2018, there has been a surge in the amount of literature related to nanophotonic inverse design with deep learning, apart from a few efforts [36] prior to this period.

The ways in which deep neural networks have been exploited for inverse design are multifaceted. However, the basic idea boils down to training a neural network to learn an approximation $f_{approx}$ of the function $f$ that maps a design $D$ in the design space $\mathcal{D}$ to the corresponding EM response $R$ in the space of responses $\mathcal{R}$ (Fig. 13). Although preparing the necessary data for training and the training itself might be arduous and time-consuming, the training is only a one-time effort. Once such a network is trained, it can be used to accelerate the forward computation/simulation for the candidate designs within an iteration of conventional iterative approaches. Such approximate models that take the place of lengthy and computation-intensive forward simulations are known as "surrogate models" or "metamodels." Deep learning also enables a whole new paradigm of inverse design, where a neural network is trained to learn the inverse mapping from $\mathcal{R}$ to $\mathcal{D}$, and then it is deployed to predict the geometry $D$ for a desired response $R$ directly without having to undergo the iterative optimization routine.

#### *4.1c. Neural Networks as Surrogate Models*

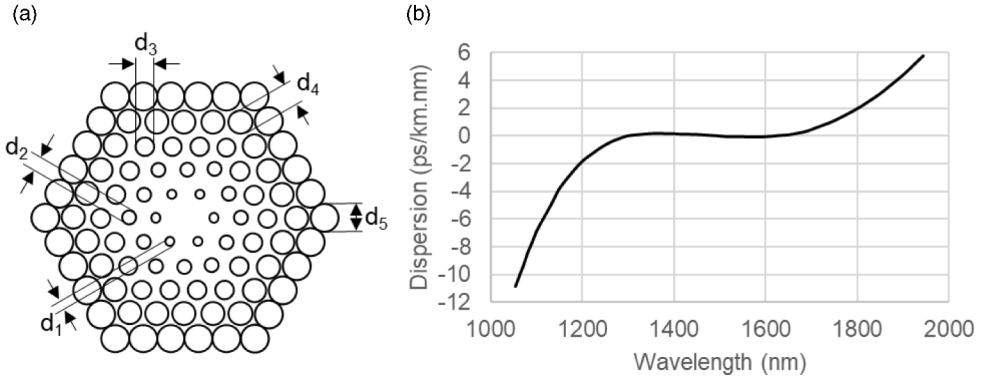
A straightforward way deep learning is exploited for nanophotonic inverse design is by using trained deep neural networks as a surrogate (metamodel) for forward EM computation within conventional inverse design approaches outlined in Section 4.1.1. Such adaptation allows for a significant reduction in the time required for iterative inverse design because running a neural network for computing the response $R$ corresponding to a design $D$ is much faster and simpler than rigorously solving Maxwell's equations with an EM solver.

Figure 13




**Neural-network-based solutions to nanophotonic inverse problems.** (a) A neural network that is trained to approximate the forward mapping $f$ from designs $D$ to their responses $R$ can be used to evade the slow and computationally intensive step of rigorously solving Maxwell's equations for the candidate designs for an inverse design iteration. (b) A neural network can also be trained to directly learn the inverse mapping, i.e., the mapping from electromagnetic responses to corresponding designs, so that it can be used to output the design $D_0$ corresponding to a target response $R_0$ non-iteratively in just one forward pass through the network. NN, neural network.

Figure 14




**Design of an ultra-flat and zero dispersion PCF with deep learning.** (a) Five-ring PCF geometry optimized by El-Mosalmy *et al.* [36] to inverse design an ultra-flat dispersion curve. The fibers along the $i$th ring have diameter $d_i$. (b) Ultra-flat dispersion attained between 1.3 μm and 1.6 μm, by optimizing the fiber diameters within the PCF geometry. The optimization was performed by "grid search," facilitated by evaluating numerous designs very fast by a "forward" neural network, which was trained to map fiber diameters to corresponding dispersion value at various wavelengths.

To outline the integration of neural networks as a metamodel within conventional approaches, here we briefly describe the work of El-Mosalmy *et al.* [36], to design a polarization rotator and an ultra-flat, zero dispersion photonic crystal fiber (PCF) by using deep neural networks. To obtain ultra-flat and zero dispersion (Fig. 14), they optimized the air hole diameters within five rings of a PCF structure where the air

hole diameters in the first, second, third, fourth, and fifth concentric rings are denoted as $d_1$, $d_2$, $d_3$, $d_4$, and $d_5$. Taking the hole pitch as 1.7825 μm and fixing $d_4$ and $d_5$ to 1.0158 and 1.6769 μm, respectively, they computed the effective index $n_{eff}$ of the PCF structures for different values of $d_1$, $d_2$, and $d_3$ at different wavelengths (λ) using simulations based on the full vectorial finite difference method (FVFDM). Then they used the calculated data to train a neural network to predict $n_{eff}$ for values of λ, $d_1$, $d_2$, and $d_3$ that were not used not in the training data. The trained NN is then used to predict $n_{eff}$ for a large number of (λ, $d_1$, $d_2$, $d_3$) combinations on a finely spaced design grid without any extra EM simulation. They ultimately found out that ultra-flattened zero dispersion can be approximately obtained over wavelengths ranging from 1.5 μm to 1.6 μm at $d_1 = 0.53$ μm, $d_2 = 0.65$ μm, and $d_3 = 0.73$ μm.

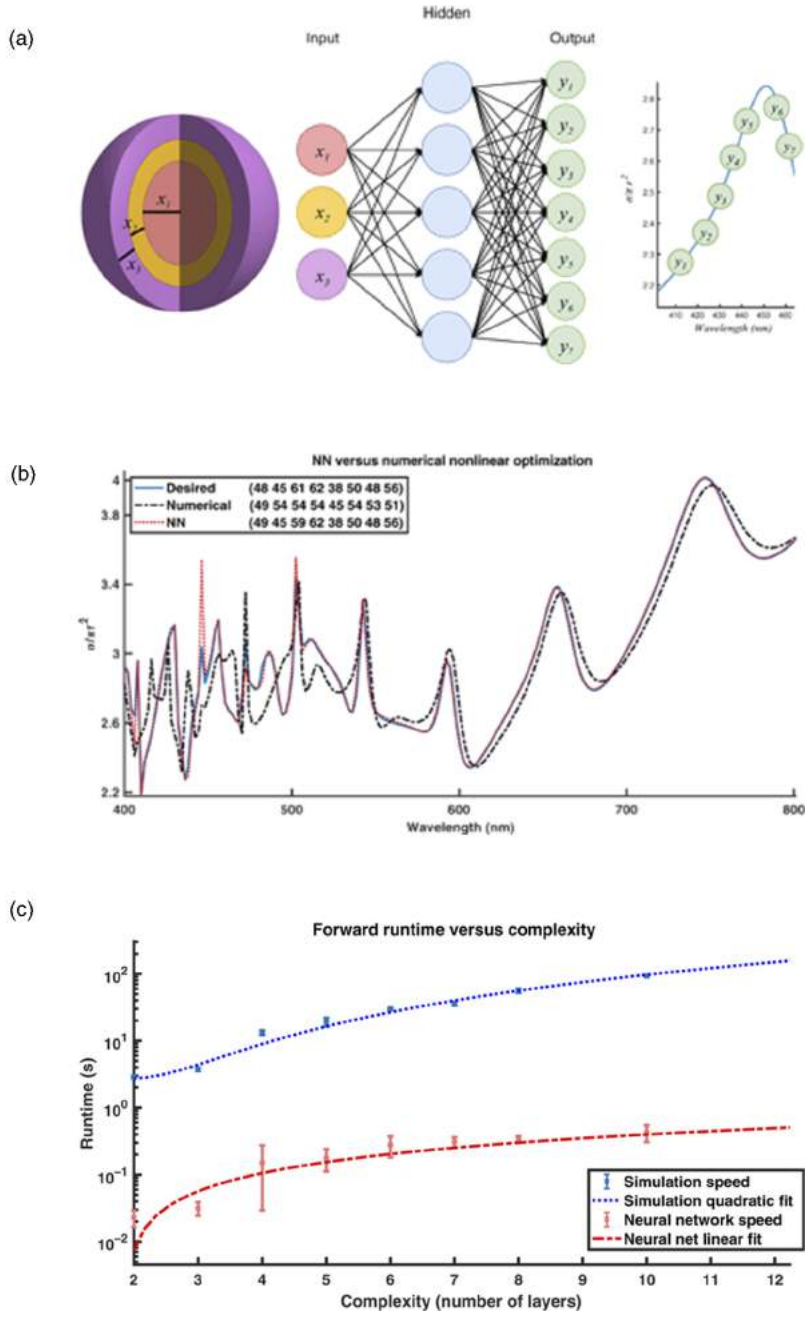
The differentiability of ANNs and the availability of efficient algorithms such as error backpropagation for computing gradients with respect to the input design parameters permit the application of powerful optimization tools such as gradient-descent method within the framework of inverse design with ANNs used as surrogate models. For example, Peurifoy *et al*. [37] trained an ANN meta-model for scattering cross section spectrum of a core–shell nanostructure as a function of the shell thickness and applied the gradient descent method to optimize the shell thicknesses to achieve a target spectral response. The retrieved values of the shell thicknesses for the complex target spectral response (Fig. 15) show a remarkable agreement with the ground truth values, demonstrating the ability of deep neural networks to model complex physical interactions and design rules through training. The number of training samples was also small, i.e., equivalent to sampling each design parameter (shell thickness) only four times. The results presented in this work summarize many of the potential incentives for incorporating deep learning into nanophotonic inverse designs.

#### *4.1d. Neural Networks for Inverse Mapping in Nanophotonics*

Apart from being used as a metamodel in conventional inverse design approaches, deep learning unlocks a new paradigm for inverse design. Instead of learning the (forward) mapping from the design parameters to the corresponding EM response, a deep neural network can also be trained to learn the (inverse) mapping from an EM response to the corresponding design structure/geometry, and as a result of this, the entire iterative optimization routine can be sidestepped, and inverse design can be obtained by a single forward pass through the inverse mapping network. However, training of such an inverse mapping neural network might be challenging because the mapping from response to geometry/structure is not one-to-one, i.e., there may exist more than one structures/designs that give rise to very similar responses (Fig. 12(a)). This fundamental property of non-uniqueness in inverse scattering problems causes naïve neural network training procedures to fail in convergence and generalization.

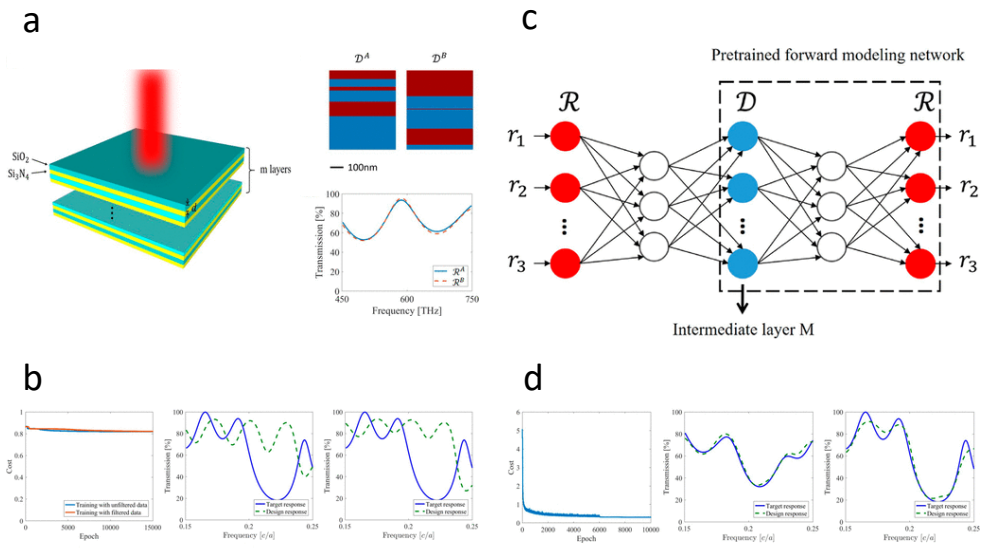
One approach to overcome the non-uniqueness challenge of inverse design solutions is the tandem network approach, which was reported by Liu *et al.* [38]. They aimed to inverse design a target transmission spectrum with a multilayer thin film that is composed of alternating layers of $SiO_2$ and $Si_3N_4$, with the thicknesses of the layers selected as the design parameters. As shown in Fig. 16(a), in the training data, there were two six-layer thin film designs with very similar transmission spectra. In the presence of such degenerate examples in the training data, the training fails to converge, as shown in Fig. 16(b). Even filtering out such examples does not improve the convergence. To circumvent this issue, the authors proposed a tandem architecture where the output of the inverse network is fed to a forward modeling network pretrained to predict the response of a design, see Fig. 16(c). Training of the inverse network is performed by minimizing the difference between the target response input to the inverse network and the response predicted by the forward modeling network for the

Figure 15




**Exploiting the differentiability of ANNs as surrogate models.** (a) Core and shell nanostructure optimized by Peurifoy *et al*. [37] to inverse design a desired scattering cross section spectrum. An ANN forward model was trained to predict the spectrum for a design parameterized by the shell thicknesses $x_i$ of the nanostructure. Then, the trained ANN was used to optimize the input shell thicknesses for a desired spectrum, by freezing the hidden layers and backpropagating to the input layer the error between the desired output and the predicted output. (b) The response of the neural network design agrees very well with the target, whereas that of a design obtained by nonlinear optimization shows significant deviations. (c) Comparison of runtime between numerical simulation and neural network forward model. Forward runtime of the trained models varies linearly with design complexity whereas the runtime of numerical simulations varies quadratically. From Peurifoy *et al*., Sci. Adv. **4**, eaar4206 (2018) [37]. Reprinted with permission from AAAS.

Figure 16




**Tandem network approach for solving non-uniqueness problem.** (a) The non-uniqueness of inverse mapping exemplified by two different six-layer thin-film designs having very similar transmission spectra. (b) Such non-uniqueness in inverse mapping causes naïve training of inverse mapping network to fail to converge even if the ambiguous examples are filtered out from the training dataset, as the responses of the network-predicted designs fluctuate significantly from the target responses for training with both the unfiltered and the filtered datasets. (c) Training of inverse mapping network in tandem configuration with a pretrained forward modeling network, as proposed by Liu *et al.* [38]. The tandem network approach works because even if the predicted design for a target response is not similar to the ground truth design, the error will be low as long as the response of the predicted design is similar to the target response. (d) Successful convergence of training of inverse mapping network in tandem configuration. After training, the inverse mapping network predicts designs with responses very similar to the target responses. Reprinted with permission from Liu *et al.*, ACS Photonics **5**, 1365 (2018) [38]. Copyright 2018 American Chemical Society.

inverse network output design. This tandem architecture partially avoids the issue of non-uniqueness because for the training cost function to be low, the designs by the inverse neural network are not required to be the same as the designs in the training data; as long as the predicted designs and the ground truth designs have similar output responses (from the forward modeling network) the overall loss function will be reduced, helping the learning and generalization of the network. The significant improvement in convergence following the training of the inverse network within the tandem configuration is evident from Fig. 16(d). The trained inverse network was also shown to predict the design parameters corresponding to arbitrarily-defined Gaussian-shaped transmission spectra within a fraction of a second, demonstrating the generalizability of the trained network and much faster inverse design capability compared with alternative iterative optimization routines.

Another approach toward mitigating the non-uniqueness issue is to use probabilistic deep learning models instead of deterministic models. Instead of predicting a solution, such models predict the statistical distribution over probable solutions, sampling from which would yield the ultimate solution. The advantage of such modeling over the tandem network approach is the ability to retrieve multiple solutions, over which

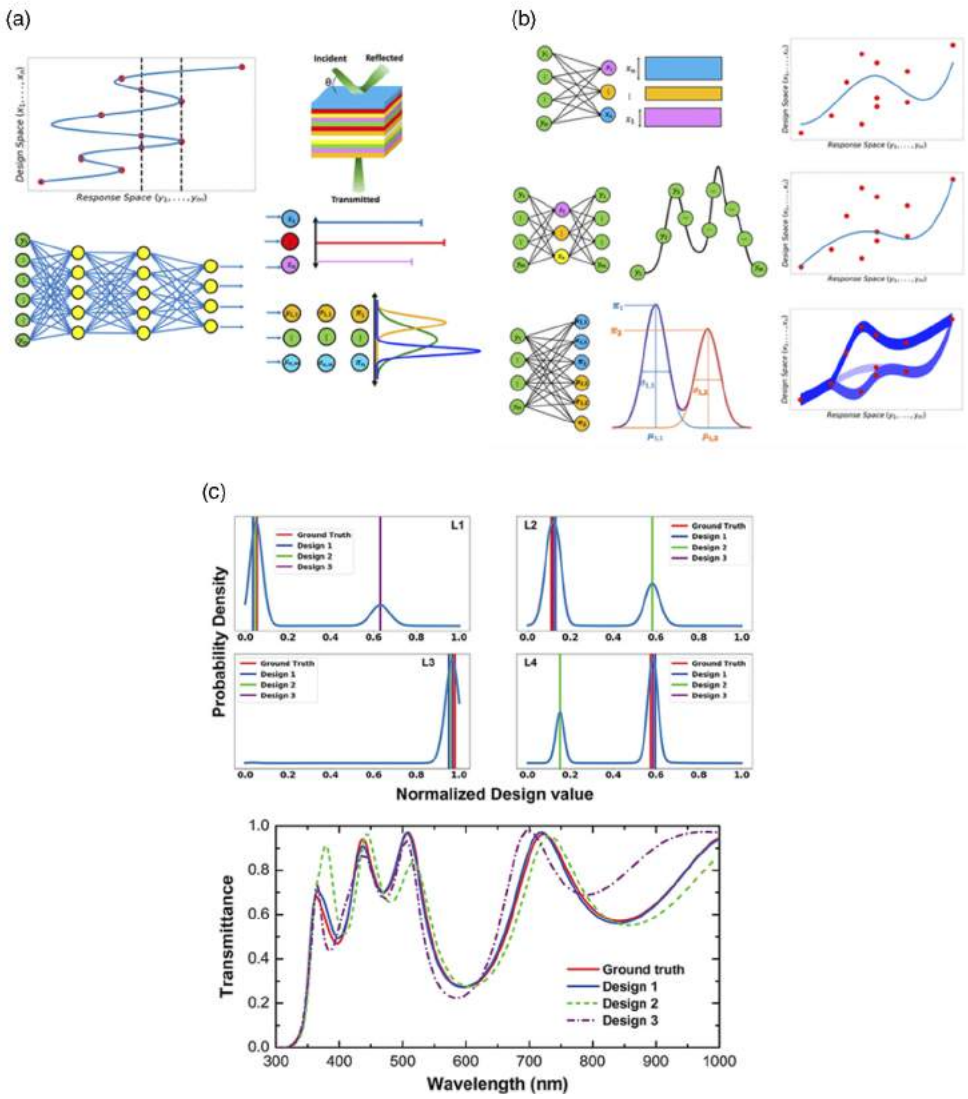
further optimization can be performed to increase the probability of reaching a better approximation. In addition, there is no need for a pretrained forward modeling network as required in the tandem network approach. For example, Unni *et al.* [301] introduced a mixture density network (MDN) approach as an alternative to tandem neural networks to address the non-uniqueness of inverse design solutions. This MDN models the design variables as having multimodal probability distributions parameterized by deterministic functions of the target response. Figure 17(b) compares the MDN approach against the tandem network approach: although the standard neural network fails to converge to any of the optimum designs, the tandem network converges to one of the ground truth designs ignoring the others and the converged solution might not be globally optimum. However, the probabilistic modeling of the design variables in MDN approach allows for the retrieval of multiple solutions through a sampling of the learned distribution, on which postprocessing was performed to obtain further refinement of these probabilistic solutions. Figure 17(c) depicts three different four-layer thin-film designs sampled from the learned parameter distributions, which successfully provide a close match to the target spectra.

Probabilistic deep generative models such as variational autoencoders (VAEs) and generative adversarial networks (GANs) can also partially lift the restriction of the design space and be used to generate free-form design geometries. Such models are generally more complex and have sub-components (encoder and decoder for VAE, generator and discriminator for GAN) that are jointly trained. Many works in the literature have used GANs for nanophotonic inverse design. For example, the work of So and Rho [302] used a conditional deep convolutional GAN to inverse design desired reflection spectra using metasurfaces. Metamaterials (and their 2D counterpart, metasurfaces) can be assembled from periodic arrays of "meta-atoms," i.e., unit cells comprising subwavelength structures made out of, e.g., metals or dielectrics. The effective optical properties of the metamaterial depend on the meta-atom constituents as well as their precise geometry (e.g., shape, size, orientation). The meta-atom architecture that was optimized in Ref. [302] is composed of a 30-nm-thick antenna on 50-nm $MgF_2$ spacer and a 200-nm silver reflector on a silicon substrate, with a unit cell dimension of 500 nm (Fig. 18(a)). A conditional deep convolutional GAN (cDCGAN) was used to generate free-form meta-atom geometries corresponding to the desired spectrum, which was provided as the condition vector. The trained cDCGAN was able to predict nanophotonic geometries closely matching the ground truth for reflection spectra not presented during training; see Fig. 18(b). The cDCGAN was also tested with completely new geometries of triangular and star-shaped antennae which did not exist in the training and validation datasets, and the results reveal that it can generalize well to spectral responses corresponding to designs of unseen shapes (see Fig. 18(c)). The generated meta-atom geometries are different from the ground truths, but the resulting reflection spectra are similar to the target reflection spectra, which is another manifestation of the non-uniqueness of the inverse mapping, i.e., different designs can have very similar responses. The trained cDCGAN was further tested with randomly generated, hand-drawn spectra with Lorentzian-like shapes, the results of which are shown in Fig. 18(d). The responses (reflection spectra) corresponding to the generated nanophotonic geometries show reasonably good agreement with the target responses. All these results confirm the generalization success of properly trained ANNs for providing fast and non-iterative inverse design solutions.

#### *4.1e. Emerging Approaches and Methods*

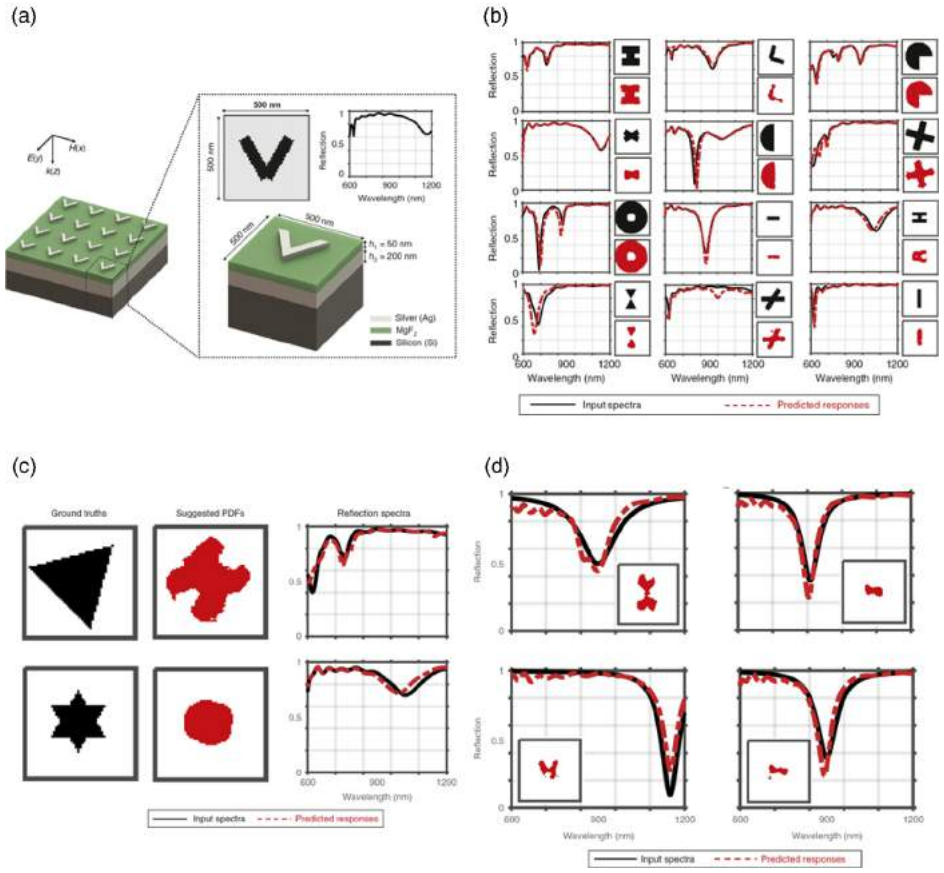
This subsection highlights some other emerging deep-learning-based approaches in photonic/optical inverse design that are distinct from the approaches described in earlier subsections. The underlying theme of a majority of these works is engineering

Figure 17




**Probabilistic modeling as a solution to non-uniqueness of inverse mapping.** (a) In the context of inverse mapping, a deterministic neural network tries to model the exact values $(x_1, \cdots, x_n)$ of the design variables, whereas a probabilistic one, such as mixture density network (MDN), models the probability distributions of design variables, parameterized by, for example, a mean $\mu$, a standard deviation $\sigma$, and a (relative) weight $\pi$ for each mode of a multimode Gaussian distribution for each design variables [301]. (b) Because of non-uniqueness of inverse mapping, a deterministic inverse network trained in isolation is highly likely to converge to a solution that does not coincide with any of the true solutions for a target response. Although a deterministic model trained in tandem configuration (also see Fig. 16) is able to predict a true solution, it ignores all the other solutions. Probabilistic MDN models, on the other hand, allow the retrieval of different solutions through sampling the learned distribution of design parameters. (c) Three different four-layer thin-film designs whose transmittance spectra closely agree with the ground truth spectrum (target), obtained by sampling the learned distributions for the film thicknesses by an MDN. Reprinted with permission from Unni *et al.*, ACS Photonics **7**, 2703 (2020) [301]. Copyright 2020 American Chemical Society.

Figure 18




**Deep generative models for inverse design of free-form shapes.** (a) Base meta-atom geometry optimized by So and Rho [302] for inverse designing target reflection spectra. They collected a dataset consisting of pairs of cross-sectional image of silver nanoantenna of either of six representative shapes (circle, square, cross, bowtie, H-shaped, and V-shaped) and corresponding reflection spectrum, and trained a cDCGAN to predict a nanoantenna shape given a target spectrum as input. (b) The evaluation of the trained cDCGAN on test data not used in training. The predicted geometry and corresponding response agrees closely with both the design and response of the ground truth in the dataset. (c) Same as (b). In these cases, however, the predicted geometry did not match with the ground truth (corresponding geometry in the dataset), although their spectral responses matched very well. This is a manifestation of the non-uniqueness of inverse mapping. (d) The neural network was also able to accurately predict meta-atom geometry for arbitrarily defined Lorentzian-like function not present in the dataset. Reprinted from So and Rho, Nanophotonics **8**, 1255 (2019) [302]. Copyright 2019 De Gruyter.

the spectral response of photonic structures such as metamaterials and metasurfaces [302–312], layered photonic structures [38,301], and core-and-shell nanoparticles [37,313]. Several of these works report the ability of generative deep learning models to predict, e.g., meta-atom geometries to create an arbitrarily defined, desired spectral response [302,305,310]. The approaches adopted in these works could also apply to nanophotonic designs for other applications such as sensing [304] and spectral filtering [41,309]. A significant fraction of these works are dedicated to metasurface design for applications such as beam engineering [314], inverse scattering [315,316], gradient metasurfaces [317–319], metagratings [320,321], among others. Deep learning

approaches have also found wide applications in inverse design for holography [322–324], color engineering [325–328], solar cell/photovoltaics [329–334], and integrated photonics [335–339].

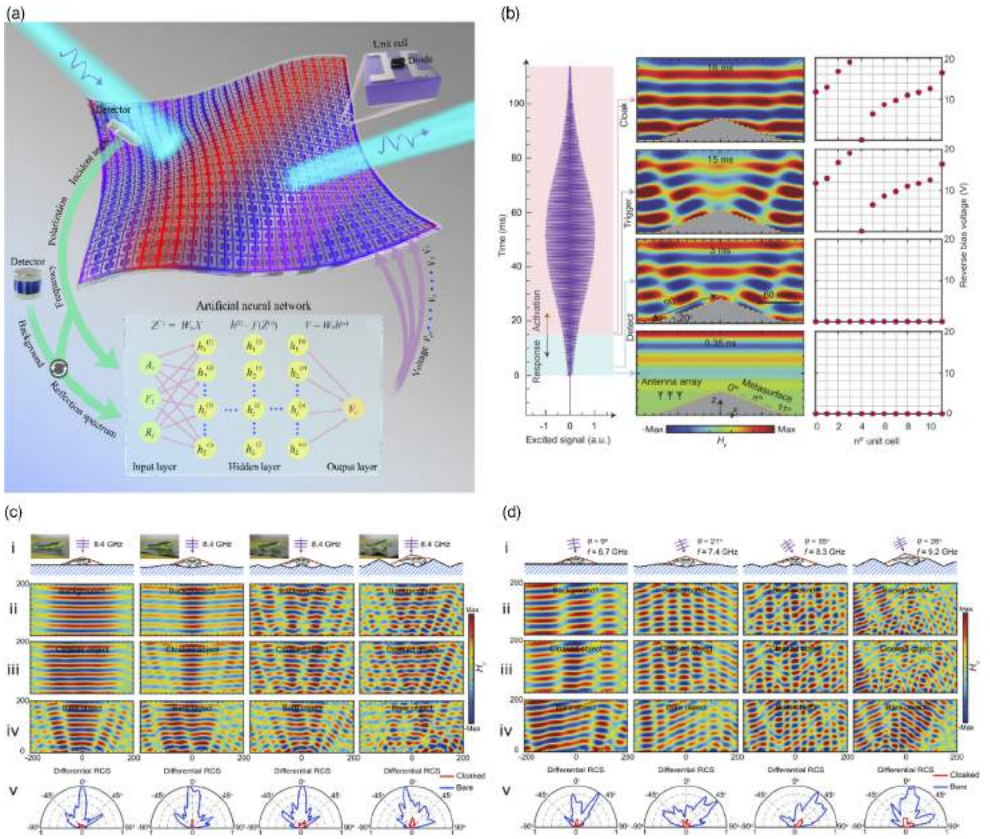
Another recent work demonstrating the use of deep learning in inverse design for an exciting application is reported by Qian *et al*. [315] where they demonstrated a self-adaptive metasurface cloak that imparts invisibility to an object and responds within milliseconds to changing incident waves and surroundings without any human intervention, see Fig. 19. This adaptive metasurface cloak consists of five main parts: a reconfigurable metasurface inclusion, two detectors, a pretrained ANN, and a power supply. The active meta-atoms of the ultrathin metasurface provide different local reflections to produce a back-scattered wave similar to that produced by the bare surrounding. The two detectors are used to probe the incident wave and the surround. The reconfigurability of the meta-atoms is attained by using loaded varactor diodes, whose capacitance can be tuned by tuning the voltage supplied by the power supply to achieve the required reflection phase. Deep learning was used in the training of a neural network to predict the required dc bias voltage for a desired local reflection phase, which, in turn, is derived from the surrounding background and incident wave. This self-adaptive invisibility cloak was demonstrated experimentally to operate in the microwave regime, and might be scaled to operate at higher frequencies.

Another emerging approach in the inverse design of photonic hardware is based on the diffractive optical network framework. In addition to their utilization for all-optical statistical inference systems, diffractive optical networks have also been used in the design of task-specific broadband photonic systems. Diffractive optical networks were designed using deep learning tools for various applications such as pulse shaping, single and dual passband spectral filtering, spatial demultiplexing of broadband radiation, and were experimentally demonstrated in the terahertz part of the spectrum [41,42]. For example, Veli *et al*. [42] reported the synthesis of an arbitrary temporal wavefront, e.g., a 15.5 ps square pulse, by processing the spectrum carried by the input terahertz pulse with a passive diffractive optical network composed of four trained phase-only diffractive layers (Fig. 20(b)). These diffractive pulse shaping networks are also tunable in the sense that the output pulse width can be adjusted by changing the axial distance between the fabricated layers. Using the diffractive network framework, Luo et al. [41] also reported the design of single and dual passband spectral filters as well as a spatially controlled wavelength demultiplexer, see Fig. 20(c) and (d). These works exemplify the potency of deep learning for the inverse design of compact and non-intuitive optical systems for specific tasks that traditionally require intricate optical designs and setup.

### 4.2. Deep-Learning-Enabled Design for Computational Imaging and Sensing

The integration of computation within the imaging framework, enabled by advances in optoelectronic sensor technologies over the past decades, has facilitated great progress in the field of computational imaging and sensing. In a computational imager or sensor [340], optics constitute only the front-end that encodes the desired optical signal, and an electronic back-end is used for decoding and further enhancing these signals following the optical (analog) to electronic (digital) conversion by optoelectronic sensors. Computational imaging, in principle, provides significant improvements over standalone optical imagers by preserving the information in the measured signal that would otherwise be lost without the encoding. The usual approach for designing such computational imaging systems has been to optimize the optics and the back-end electronics separately for their intended purpose. However, the design space that could be reached by joint optimization of the optics and the electronics [3], harnessing their synergy more effectively for better overall performance, remains relatively unexplored.

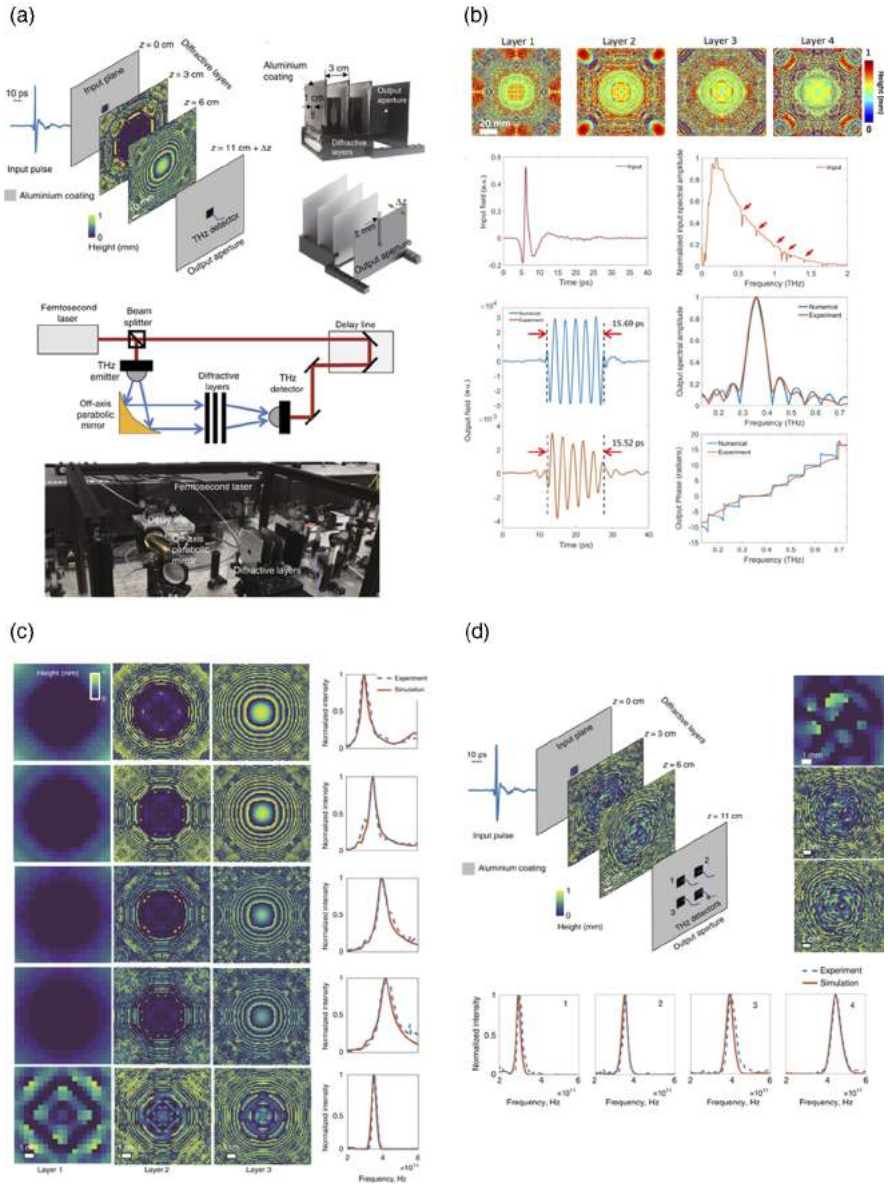
Figure 19




**Deep learning enabled design of a self-adaptive metasurface cloak.** (a) Schematic of an ultrathin layer of active meta-atoms, each incorporating a varactor diode that is independently controlled by a DC bias voltage, constituting an intelligent self-adaptive metasurface cloak [315]. In response to the incident wave and the background wave detected by the two detectors, an embedded pretrained ANN calculates all the necessary bias voltages ($V_1, V_2, \ldots, V_M$), which are then supplied to the varactor diodes to trigger cloaking. (b) Finite difference time domain (FDTD) simulation results for transient response of the cloak, where a Gaussian pulse impinging on a triangular perfect electrical conductor (PEC) bump is detected by an antenna array and then fed into a pretrained ANN, together with a hypothetically known background. The metasurface cloak is triggered, and subsequently renders the bump invisible within 15 ms. (c) Self-adaptive cloak response to four random backgrounds (row i) for normal wave incidence at 8.4 GHz in terms of near-field magnetic field distributions of the background (row ii), cloaked object (row iii) and bare object (row iv); and far-field differential radar cross section (RCS) of the cloaked (red) and bare (blue) objects for the four cases (row v). (d) Same as (c), but for random and simultaneous changes in both the incident wave (angle and frequency) and the background. Reprinted by permission from Macmillan Publishers Ltd: Qian *et al.*, Nat. Photonics **14**, 383 (2020) [315]. Copyright 2020.

Deep learning equips us with the ability to better explore this unique and timely opportunity by enabling task-specific, joint optimization of the optics and the electronics. It is possible to form a learning model by cascading an accurate numerical model of the front-end optics and the digital reconstruction/processing model (back-end), and to perform end-to-end training of the cascaded model and optimize the parameters of interest for a given desired task, as long as the models for both the optical

Figure 20




**Inverse design of broadband photonic systems using the D$^2$NN framework.** (a) Diffractive optical networks, in general, comprise spatially engineered diffractive surfaces that are trained through deep learning inside a computer. The forward training model of diffractive networks can be constructed to reflect material dispersion enabling these platforms to offer task-specific solutions for challenging broadband optical inverse design problems e.g., lensless pulse shaping [42] and spatially controlled spectral demultiplexing [41]. (b) Processing of an input terahertz pulse to optically synthesize a rectangular pulse of desired width using a four-layer diffractive optical network design [42]. Reprinted by permission from Macmillan Publishers Ltd: Veli *et al*., Nat. Commun. **12**, 37 (2021) [42]. Copyright 2021. (c) Spectral filtering of broadband terahertz pulses using three-layer diffractive optical networks [41]. (d) Spatially controlled demultiplexing of the spectral components in a broadband terahertz pulse with a two-layer diffractive optical network [41]. (a), (c), and (d) Reprinted by permission from Macmillan Publishers Ltd: Luo *et al*., Light. Sci. Applicat. **8**, 112 (2019) [41]. Copyright 2019.

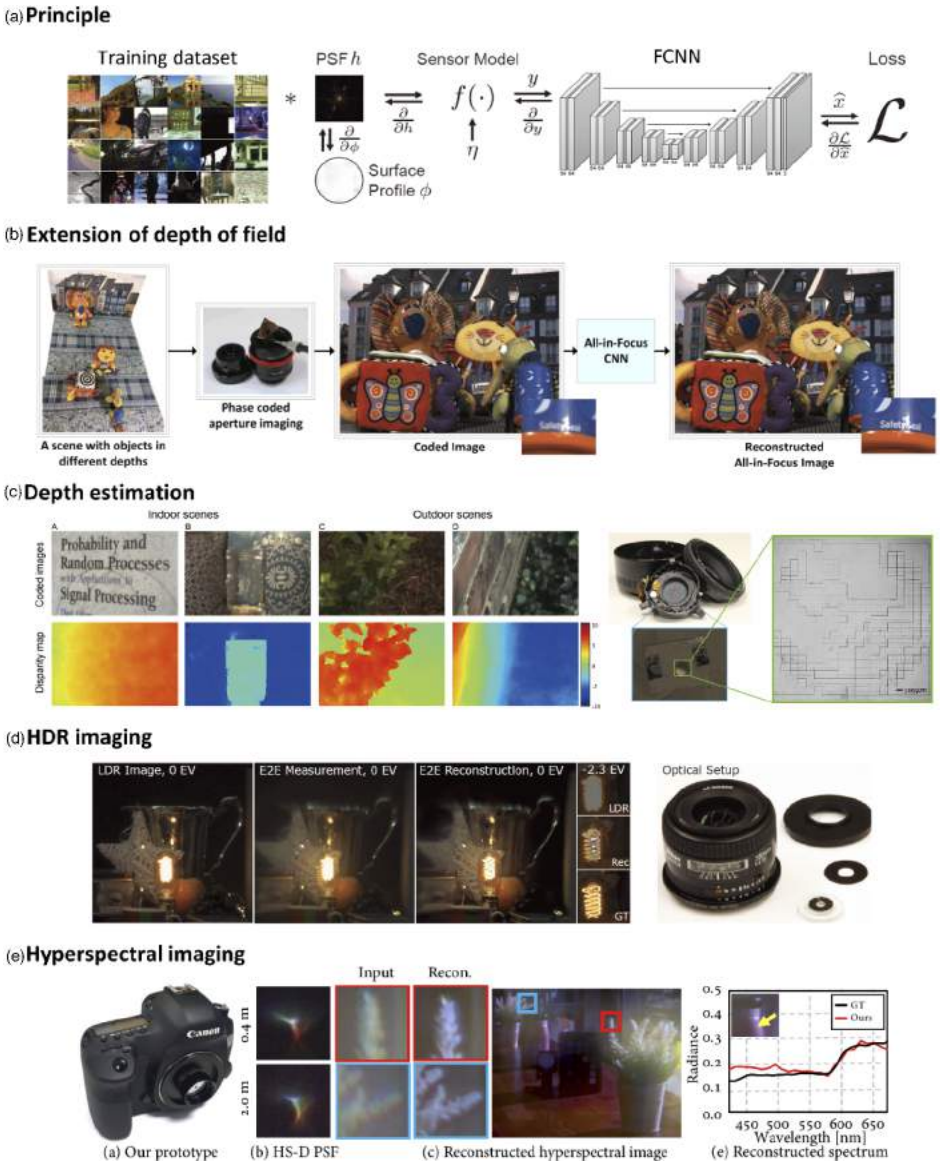
front-end and the digital back-end are differentiable with respect to these parameters to be optimized. We elaborate on this approach and the underlying opportunities in Sections 4.2.1 and 4.2.2. Apart from this end-to-end training based on explicit forward modeling of the optics, deep learning can also be used to implicitly learn the inverse mapping from measurements to design parameters for adaptive control of the desired optical system. The instances of performing inverse mapping for computational imaging and sensing is elaborated on in Section 4.2.3. We should clarify that methods and applications where deep learning is only used to optimize the back-end electronic hardware [341–343] or purely software-based digital processing models [6–9,11,13,14,17–20,22–25,31–33,344–347] used for, e.g., image reconstruction or super-resolution, lie outside the scope of this review, although they also offer various solutions to computational imaging and sensing problems in the digital domain.

#### *4.2a. End-to-End Optimization of PSF and Deep Image Reconstruction Models*

One of the ways computational imaging bypasses the limitations of standalone optics is through PSF engineering, i.e., by engineering the spatial transmittance of the pupil plane of the imaging system. The idea behind PSF engineering is to encode the optical information in such a way that information otherwise unresolvable by standalone optics with a regular PSF (e.g., depth information) can be resolved with the help of subsequent digital decoding. PSF engineering, also known as coded aperture imaging, has been widely applied in the design of computational cameras and microscopes [348–352]. In recent years, the exploitation of deep learning for jointly optimizing the PSF and the digital decoder has provided new and rich opportunities for exploration of the design space of PSFs, and yielded coded aperture designs that are high performance, task specific, but non-intuitive.

In practice, PSF engineering can be realized by introducing a coded aperture at the pupil plane, for example, a phase mask with an optimized surface profile $\phi$. The modulation of the light field by the coded aperture is equivalent to performing a convolution at the image plane, the kernel for which is the same as the PSF $h$ (Fig. 21(a)). Following an imaging measurement by the sensor, an electronic decoder is used to undo the encoding performed by the coded aperture. This decoding can be performed by a trained CNN; for example, a fully convolutional neural network (FCNN), as in Fig. 19(a), is often used to recover the target image with the desired information from the measurement. FCNN takes its name from being composed of purely convolutional up-/down-sampling layers. Given that the mapping from the surface profile $\phi$ to the PSF $h$ and a differentiable numerical model of the optical measurement $f(\cdot)$ are available, both the surface profile $\phi$ and the FCNN can be jointly optimized by backpropagation of gradients of the loss $\mathcal{L}$ between the network's image reconstruction and the target images. After convergence to a desired solution, the optimized surface profile can be conveniently fabricated, e.g., through 3D printing or lithography to physically perform the desired PSF. In recent years, the computational imaging community has developed a variety of computational cameras and displays with specific functions based on this concept of end-to-end training. One of the earlier demonstrations comes from the work by Elmalem *et al.* [353], who used a single-ring phase pattern with the radius and the phase delay of the ring optimized jointly with a CNN to achieve an extended DOF; see Fig. 21(b). Following a similar principle, Akpinar *et al.* [354] used a DOE in a coded aperture imaging system to extend the imaging DOF. In addition to the extension of the imaging DOF, researchers also demonstrated holistically optimized computational cameras using phase-coded apertures for other tasks, such as image super-resolution [43,355], monocular depth estimation [356–358], high-dynamic-range imaging [44,359], hyperspectral imaging [360,361], and light field sensing [362]; see Fig. 21(c), (d), and (e). In addition to

Figure 21




**Deep-learning-enabled computational camera designs based on PSF engineering.** (a) The general principle of PSF-engineered computational camera designs, where the surface profile of the coded aperture of the camera can be jointly trained with the CNN-based reconstruction model at the back-end using backpropagation of gradients of the same task-specific loss function. This principle has been used to improve various performance metrics in numerous computational imaging tasks including: (b) extension of the depth of field [353], reprinted with permission from [353], copyright 2018 Optical Society of America; (c) depth estimation [357]; (d) high dynamic range imaging [44]; and (e) hyperspectral imaging [361].

camera systems, researchers also extended the idea of joint optimization to designing computational near-eye display systems [363,364].

PSF engineering has also been used for high-resolution 3D imaging in computational microscopy. Researchers have been using engineered PSFs such as double helix [350,352] and Tetrapod [365,366] functions for better estimation of depth (axial distance) of objects for more than a decade. In recent years, the use of deep learning

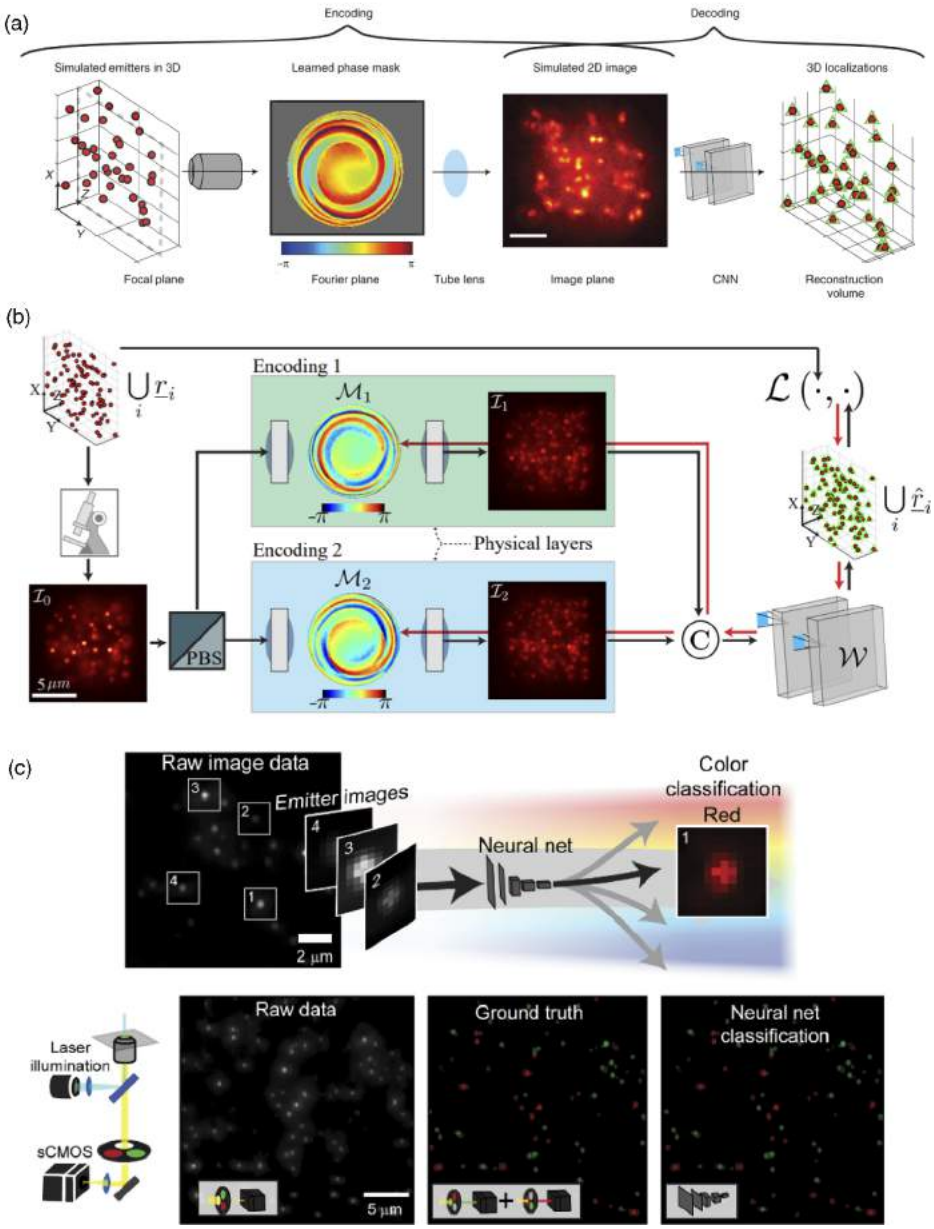
for synergistic co-optimization of optics and reconstruction algorithms for PSF engineering in computational microscopy has kickstarted. For example, Nehme *et al.* demonstrated 3D localization of dense emitters over a large axial range by using a framework named DeepSTORM3D to design an optimized PSF [10]; see Fig. 22(a). The authors placed an optimal phase mask at the Fourier plane of a 4*f* system that was used to extend the intermediate image plane formed after the tube lens. The optimal phase profile of the phase mask was obtained through a joint optimization with a CNN back-end, the purpose of which was to accurately predict the 3D positions of fluorescent emitters from the acquired 2D images. For the joint optimization, the authors used deep learning to train a model that incorporated a differentiable physical layer simulating the PSF modulation. After training their model on a large number of simulated images of randomly distributed, densely populated fluorescent emitters, the authors experimentally validated the efficacy of the framework by precisely predicting the axial positions of emitters over a depth range of ~4 μm. In a follow-up work, the authors demonstrated the prediction of the emitter positions with a 3D precision of 30 nm over an axial range of ~5 μm [367]. This enhancement was realized by introducing, instead of a single PSF, a pair of engineered PSFs along two parallel optical paths, as depicted in Fig. 22(b).

In addition to depth estimation, the idea of jointly optimizing a phase mask and a back-end CNN has also been used to perform color microscopic imaging with a monochromatic sensor, following similar ideas as in Refs. [360,369]. For example, Hershko *et al.* used deep learning to perform the automatic design of a phase mask for color encoding in localization microscopy [368]. Owing to the divergence of phase delay for different wavelengths, the designed phase mask possesses distinct PSFs for different colors, in order to encode the colors of fluorospheres into different spatial patterns. As shown in Fig. 22(c), the 2D grayscale images of these fluorosphere patterns were captured by a monochromatic camera in a standard fluorescence microscope, and the color information of the fluorospheres was recovered by a jointly trained CNN. Thus, color fluorescence microscopy imaging was achieved without any additional hardware.

#### *4.2b. End-to-End Optimization of Structured Illumination and Deep Reconstruction Models for Super-Resolution*

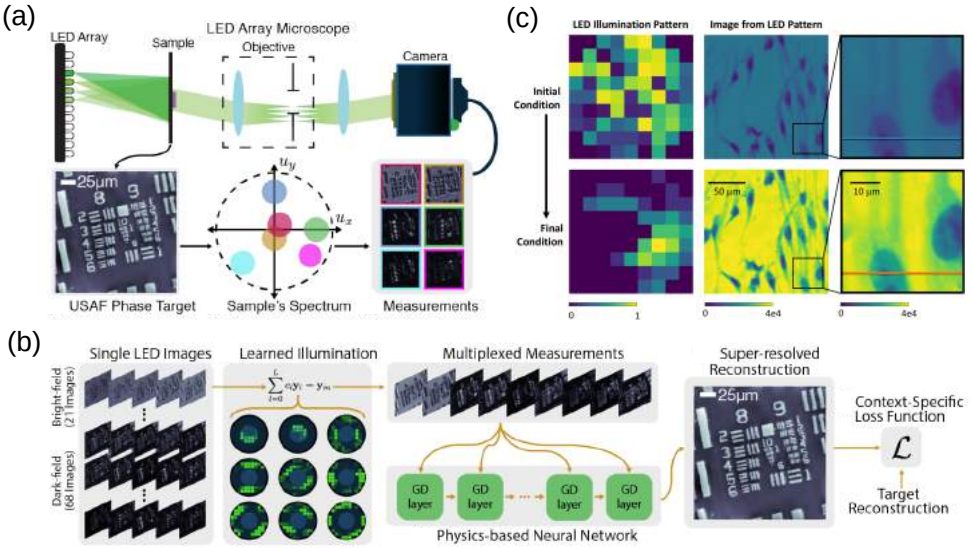
In addition to PSF engineering, manipulation of the illumination pattern, known as structured illumination, can also improve the performance of a microscope by extracting desired information from samples. The idea of structured illumination dates back to the early 2000s when Gustafsson surpassed the lateral resolution limit of light microscopy through illumination with moiré fringe patterns [370]. Fourier ptychographic microscopy is also used for super-resolution, in which multiple low-resolution images are captured with different illumination angles. Each of the low-resolution images corresponds to spatial frequency components contained in a sub-aperture with a low numerical aperture (NA) on the Fourier domain [371]. Iterative phase retrieval algorithms or alternatively, trained deep neural networks can then be used to put together these sub-apertures to reconstruct a high-resolution image of the sample. A regular implementation of multiangle illumination is based on using a 2D array of programmable light-emitting diodes (LEDs) placed before the sample, where the LEDs are switched on one by one to achieve separate and distinct angles of illumination on the sample; see Fig. 23(A). This approach of switching on a single LED for each measurement results in large time costs and hinders the imaging throughput. For solving this issue, multiplexing of LEDs that are simultaneously on for a single measurement was used to reduce the total number of measurements [372]. Subsequently, deep-learning-based joint optimization was utilized to design such multiplexed solutions, which

Figure 22




**Deep learning for PSF engineering in computational microscopy.** (a) A deep-learning-based framework for designing an optimal PSF to localize dense emitters in 3D [10]. The encoding phase-mask and the decoding CNN are jointly trained. Reprinted by permission from Macmillan Publishers Ltd: Nehme *et al.*, Nat. Meth. **17**, 734 (2020) [10]. Copyright 2020. (b) An improved version of [10], where two imaging paths with co-optimized PSFs along each are combined to enhance the performance of localization in high-density volumetric samples [367]. © 2021 IEEE. Reprinted, with permission, from Nehme *et al.*, IEEE Trans. Pattern Anal. Machine Intell. **43**, 2179 (2021) [367]. (c) Deep-learning-based hardware-software co-design is used for engineering the PSF of a microscope together with the image reconstruction neural network. The presented framework reconstructs color images using solely a standard fluorescence microscope and a grayscale camera as hardware [368]. Reprinted with permission from [368]. Copyright 2019 Optical Society of America.

Figure 23




**Deep-learning-based optimization of illumination patterns/sequences in optical imaging and microscopy.** (A) Fourier ptychographic microscopy (FPM) with an LED array source, where each LED modulates a different part of the sample's Fourier components into the passband of the microscope, increasing the spatial resolution of the images [375]. In the earlier approaches to FPM, the LEDs were turned on one by one for successive frames during imaging, so the high spatial resolution was accompanied with low temporal resolution. (B) Deep learning is used to optimize a reduced number of illumination patterns in FPM, for imaging with more optimal trade-off between temporal and spatial resolution [375]. A physics-based unrolled neural network is used to reconstruct a super-resolved image from multiplexed measurements using a set of learned LED patterns. (C) Deep-learning-enabled single-shot FPM by further reducing the number of illumination patterns in LED array-based microscopy to one [373]. Reprinted with permission from [373]. Copyright 2019 Optical Society of America.

led to a single-shot Fourier ptychographic microscopy [373,374]. In other related work, Kellman *et al.* used a physics-based learning approach to reduce the number of illuminations in LED array-based microscopy, achieving amplitude and QPI for context-specific applications with a more favorable trade-off between temporal resolution and image reconstruction quality [375,376]; see Fig. 23(B). An important difference in their approach from the others is the use of a physics-based unrolled neural network to iteratively reconstruct the high-resolution sample image. The inclusion of this optimization approach provides robustness and generalizability in addition to reducing the number of learnable parameters. In another implementation, Robey *et al.* [374] built a differentiable numerical simulation layer, which simulates the physics of low-resolution image formation of a sample under illumination with multiple LEDs. Feeding the simulated low-resolution images to a CNN that is jointly trained with the LED patterns, they were able to demonstrate single-shot imaging with improved temporal resolution and space-bandwidth product; see Fig. 23(C).

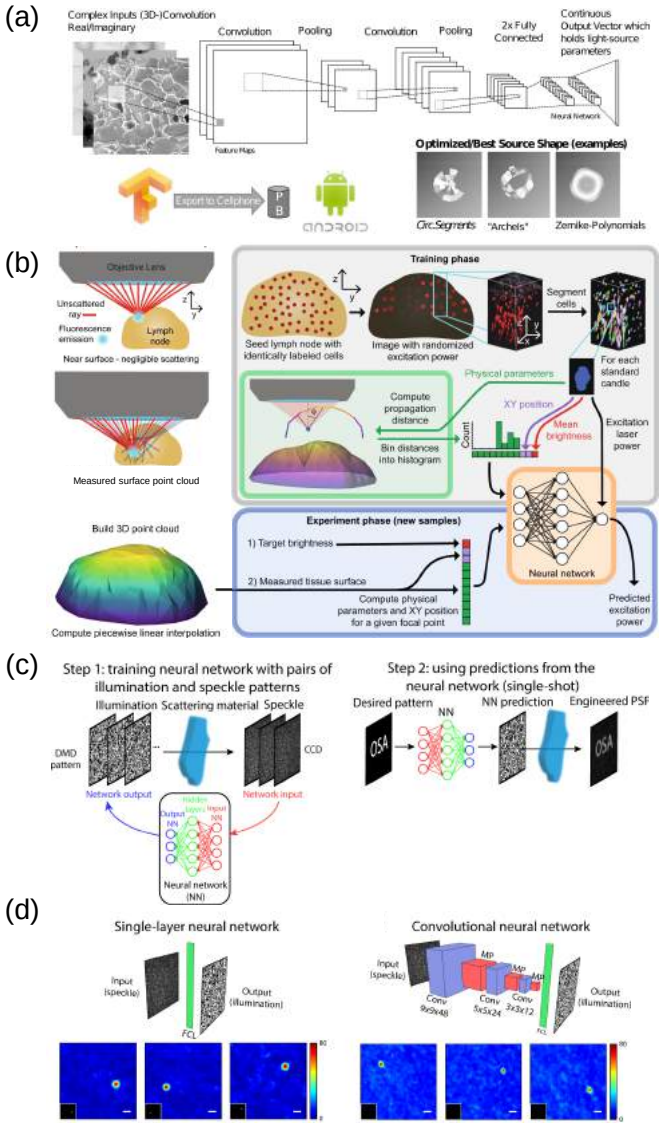
#### *4.2c. Deep Learning for Inverse Mapping in Computational Imaging and Sensing*

In Sections 4.2.1 and 4.2.2, we discussed the usage of end-to-end deep learning to holistically optimize the front-end optics and its back-end digital reconstruction

models, where the optical system is physically simulated in the forward model, and its optical parameters are optimized in the design process. However, analogous to the design scheme of inverse mapping in nanophotonics that we elaborated on in Section 4.1.4, obtaining appropriate hardware parameters can also be achieved by finding a mapping function from the desired output of the optical system to the system parameters. The critical difference between these two schemes lies in that in the latter scheme, the parameters to optimize are the internal parameters of the mapping function, which can be represented as the implicit weights of a neural network using deep learning and have no explicit physical meaning. The same idea of performing this kind of inverse design of optical systems using deep learning has also been explored in microscopy and other imaging and sensing systems [340,377–379]. For example, Diederich *et al.* [380] trained a CNN that was used to tailor the shape of the light source in a smartphone microscope to the features of different transparent samples, so that the phase-contrast images of the samples can be adaptively enhanced for better visualization (see Fig. 24(a)). A similar application of deep neural network for multiphoton illumination microscopy was reported in Ref. [381], where the authors used a learning-based approach to estimate the correct illumination power in curved samples compared to a conventional multiphoton microscope, and adaptively applied the minimal illumination needed to observe the structures of interest; see Fig. 24(b). The advantages of such learning-based adaptive control of illumination intensity include affordable photon budget and reduced perturbation/toxicity to the sample induced by the imaging process [381].

Another example of performing inverse mapping using deep learning is for wavefront sensing and correction. It is an important challenge to mitigate variable distortions brought by, for example, scattering media around the sample. Adaptive optics has a rich history to mitigate such challenges and has stepped onto a new level with the help of deep learning. Neural network-based approaches were demonstrated earlier in array telescopes to predict piston and tilt errors within the array elements using a network with one hidden layer [383]. Following the progress made in deep learning in the last decade, the use of deep learning in adaptive optics has significantly increased recently. For example, researchers have demonstrated the efficacy of using deep neural networks to provide necessary wavefront corrections to shape the optical beam after a scattering medium. Specifically, they trained deep neural networks to, for example, learn the inverse mapping from speckle patterns to illumination of the scattering media, so that once the network is trained it can generate the corrected illumination corresponding to desired speckle patterns [382]; see Figs. 22(c) and (d). In efforts related to image-based wavefront sensing and correction, researchers have also used CNNs to retrieve the spatial distribution of the wavefront errors based on the measured image intensities, so that optical components such as SLMs can be used on demand to compensate for such wavefront errors. Such efforts include performing wavefront sensing based on measured PSFs [384,385], high-order aberration prediction using the pattern measured by a Shack–Hartmann wavefront sensor [45], and correction of aberrations in excitation and detection of optical microscopy based on reflected light from scattering samples [46]. All-optical solutions to inverse problems in computational imaging applications have been explored recently [324,386,387]. For example, application of deep learning to design diffractive optical networks (see Section 3.2) for twin-image free all-optical reconstruction of inline holograms was reported in Ref. [324], whereas Luo *et al.* [386] demonstrated "seeing through" random, unknown diffusers using diffractive networks that all-optically undo the scattering introduced by these diffusers. The advantages of these diffractive approaches over their digital counterparts comprise the speed of image reconstruction and the elimination of computing power (except for the illumination light) due to the passive nature of diffractive networks.

Figure 24




**Deep learning of inverse mapping in computational imaging and sensing.** (a) Enhanced phase contrast by adjusting the illumination shape adaptively to the features of the transparent samples is achieved in a smartphone microscope. The adaptive adjustment of illumination shape was enabled by incorporating a trained CNN that predicts the optimum illumination parameters for a given sample from the sample features [380]. Reprinted from Diederich *et al.*, PLoS One **13**, e0192937 (2018) [380]. (b) A neural network is used in multiphoton illumination microscopy to decide the appropriate excitation power as a function of the sample surface profile [381]. Reprinted by permission from Macmillan Publishers Ltd: Pinkard *et al.*, Nat. Commun. **12**, 1916 (2021) [381]. Copyright 2021. (c) and (d) Light control through scattering media with neural networks [382]. (c) A neural network (NN) is trained with pairs of illumination and corresponding speckle pattern through a scattering material. Once the NN is trained, it is used to predict the illumination necessary to generate a target pattern after the scattering material. (d) The ability of single-layer neural networks or CNNs trained likewise to predict necessary illumination patterns for generating diffraction-limited Gaussian foci through scattering media at different positions within the field of view. (c) and (d) Reprinted with permission from [382]. Copyright 2018 Optical Society of America.

## 5. FUTURE OUTLOOK

The field of optics and photonics has been profoundly influenced by deep learning, one of the most disruptive technologies of the last decades. Advances in deep learning research will continue to bring about innovative design approaches for optical systems and nanophotonic components. Deep learning could even potentially help fundamental knowledge discovery in some branches of optical sciences. It could also unveil unconventional, non-intuitive, and/or task-specific solutions to a wide range of inverse design problems in optics and photonics that either had not been solved previously or the conventional solutions had serious shortcomings in terms of system parameters and targeted performance metrics.

The field of optics and photonics bears the promise of enabling new computing technologies that would help deep neural networks step into their next phase of evolution in terms of speed, power efficiency, and scalability. Although it is arguable if we can realize general-purpose optical computers as a practical technology in the foreseeable future, next-generation AI applications and their requirements make optical computing an intriguing research area that is full of opportunities (and challenges). It is not possible to know in advance whether or when optics will fully deliver on the promises it bears, but it is highly likely that deep learning will benefit from the bold attempts made by photonics researchers in this regard. All-optical implementations of deep neural network equivalent processors are not yet on the horizon, one of the major obstacles being the all-optical implementation of low-power, scalable, and practical nonlinear activation functions. However, nonlinear deep neural networks with photonic accelerators may very well be in vogue within the next few years, whereas niche applications benefiting from the high speed and parallelism of optical implementations could be on the rise. It might not even be far-fetched to anticipate rises in applications of linear optical networks, for various machine vision applications that demand extreme speed, ultra-low power consumption, and ubiquitous computing.

Hybrid neural network systems present another key direction that can merge the best of both worlds, i.e., the bandwidth, speed, power efficiency of optics/photonics and the flexibility of electronic digital computing. Optical–electronic inference engines realized by the joint optimization of the two computing modalities might bring optical computing and its advantages into practical applications in, e.g., computer vision, microscopy, and robotics. It is futile to speculate on the limit optics can go, if there is one, in transforming deep learning, but the advances in recent years exhort for continuing explorations. As for researchers that work at the intersection of deep learning and optics/photonics fields, time is ripe for reaping the benefits from this symbiotic relationship and collaboration between fields, which will continue to provide exciting opportunities for the decades to come.

## FUNDING

Office of Naval Research; Air Force Office of Scientific Research.

## ACKNOWLEDGMENTS

The authors acknowledge the US Air Force Office of Scientific Research (AFOSR), Materials with Extreme Properties Program funding and the US Office of Naval Research (ONR) EO/IR Sensors and Sensor Processing Program funding.

## DISCLOSURES

The authors declare no conflicts of interest.

## DATA AVAILABILITY

No data were generated or analyzed in the presented research.

## REFERENCES


1. R. Athale and D. Psaltis, "Optical computing: past and future," Opt. Photonics News **27**, 32 (2016).
2. M. Schmeisser, B. C. Heisen, M. Luettich, B. Busche, F. Hauer, T. Koske, K.-H. Knauber, and H. Stark, "Parallel, distributed and GPU computing technologies in single-particle electron microscopy," Acta Crystallogr., Sect. D: Biol. Crystallogr. **65**, 659–671 (2009).
3. G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," Nature **588**, 39–47 (2020).
4. F. Niesler and M. Hermatschweiler, "Two-photon polymerization - a versatile microfabrication tool: from maskless lithography to 3D printing," Laser Tech. J. **12**, 44–47 (2015).
5. M. Emons, K. Obata, T. Binhammer, A. Ovsianikov, B. N. Chichkov, and U. Morgner, "Two-photon polymerization technique with sub-50 nm resolution by sub-10 fs laser pulses," Opt. Mater. Express **2**, 942 (2012).
6. Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," Optica **4**, 1437 (2017).
7. T. Nguyen, T. Nguyen, Y. Xue, Y. Li, L. Tian, L. Tian, and G. Nehmetallah, "Deep learning approach for fourier ptychography microscopy," Opt. Express **26**, 26470–26484 (2018).
8. H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydın, L. A. Bentolila, C. Kural, and A. Ozcan, "Deep learning enables cross-modality super-resolution in fluorescence microscopy," Nat. Methods **16**, 103–110 (2019).
9. Y. Wu, Y. Rivenson, H. Wang, Y. Luo, E. Ben-David, L. A. Bentolila, C. Pritz, and A. Ozcan, "Three-dimensional virtual refocusing of fluorescence microscopy images using deep learning," Nat. Methods **16**, 1323–1331 (2019).
10. E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, T. Naor, R. Orange, T. Michaeli, and Y. Shechtman, "DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning," Nat. Methods **17**, 734–740 (2020).
11. Y. Wu, Y. Luo, G. Chaudhari, Y. Rivenson, A. Calis, K. de Haan, and A. Ozcan, "Bright-field holography: cross-modality deep learning enables snapshot 3D imaging with bright-field contrast using a single hologram," Light: Sci. Appl. **8**, 25 (2019).
12. Y. Rivenson, Y. Wu, and A. Ozcan, "Deep learning in holography and coherent imaging," Light: Sci. Appl. **8**, 1–8 (2019).
13. Y. Wu, Y. Rivenson, Y. Zhang, Z. Wei, H. Günaydin, X. Lin, and A. Ozcan, "Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery," Optica **5**, 704 (2018).
14. Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," Light: Sci. Appl. **7**, 17141 (2018).
15. G. Barbastathis, A. Ozcan, and G. Situ, "On the use of deep learning for computational imaging," Optica **6**, 921 (2019).
16. Y. Rivenson, T. Liu, Z. Wei, Y. Zhang, K. de Haan, and A. Ozcan, "PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning," Light: Sci. Appl. **8**, 23 (2019).

17. N. Borhani, E. Kakkava, C. Moser, and D. Psaltis, "Learning to see through multimode fibers," Optica **5**, 960 (2018).
18. B. Rahmani, D. Loterie, G. Konstantinou, D. Psaltis, and C. Moser, "Multimode optical fiber transmission with a deep learning network," Light: Sci. Appl. **7**, 1–11 (2018).
19. M. E. Kandel, Y. R. He, Y. J. Lee, T. H.-Y. Chen, K. M. Sullivan, O. Aydin, M. T. A. Saif, H. Kong, N. Sobh, and G. Popescu, "Phase imaging with computational specificity (PICS) for measuring dry mass changes in sub-cellular compartments," Nat. Commun. **11**, 6256 (2020).
20. N. Goswami, Y. R. He, Y.-H. Deng, C. Oh, N. Sobh, E. Valera, R. Bashir, N. Ismail, H. Kong, T. H. Nguyen, C. Best-Popescu, and G. Popescu, "Label-free SARS-CoV-2 detection and classification using phase imaging with computational specificity," Light: Sci. Appl. **10**, 176 (2021).
21. Y. Park, C. Depeursinge, and G. Popescu, "Quantitative phase imaging in biomedicine," Nat. Photonics **12**, 578–589 (2018).
22. V. Bianco, P. L. Mazzeo, M. Paturzo, C. Distante, and P. Ferraro, "Deep learning assisted portable IR active imaging sensor spots and identifies live humans through fire," Opt. Lasers Eng **124**, 105818 (2020).
23. V. Bianco, P. Memmolo, P. Carcagnì, F. Merola, M. Paturzo, C. Distante, and P. Ferraro, "Microplastic identification via holographic imaging and machine learning," Adv. Intell. Syst. Comput. **2**, 1900153 (2020).
24. S. You, E. J. Chaney, H. Tu, Y. Sun, S. Sinha, and S. A. Boppart, "Label-free deep profiling of the tumor microenvironment," Cancer Res. **81**, 2534–2544 (2021).
25. S. K. Mirsky, I. Barnea, M. Levi, H. Greenspan, and N. T. Shaked, "Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning," Cytometry, Part A **91**, 893–900 (2017).
26. G. Dardikman and N. T. Shaked, "Phase unwrapping using residual neural networks," in *Imaging and Applied Optics 2018 (3D, AO, AIO, COSI, DH, IS, LACSEA, LS&C, MATH, PcAOP)* (Optical Society of America, 2018), p. CW3B.5.
27. J. Yoon, Y. Jo, M. Kim, K. Kim, S. Lee, S.-J. Kang, and Y. Park, "Identification of non-activated lymphocytes using three-dimensional refractive index tomography and machine learning," Sci. Rep. **7**, 6654 (2017).
28. Y. Jo, S. Park, J. Jung, J. Yoon, H. Joo, M. Kim, S.-J. Kang, M. C. Choi, S. Y. Lee, and Y. Park, "Holographic deep learning for rapid optical screening of anthrax spores," Sci. Adv. **3**, e1700606 (2017).
29. J. Li, J. Garfinkel, X. Zhang, D. Wu, Y. Zhang, K. de Haan, H. Wang, T. Liu, B. Bai, Y. Rivenson, G. Rubinstein, P. O. Scumpia, and A. Ozcan, "Biopsy-free in vivo virtual histology of skin using deep learning," Light: Sci. Appl. **10**, 233 (2021).
30. Y. Rivenson, H. Wang, Z. Wei, K. de Haan, Y. Zhang, Y. Wu, H. Günaydın, J. E. Zuckerman, T. Chong, A. E. Sisk, L. M. Westbrook, W. D. Wallace, and A. Ozcan, "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning," Nat. Biomed. Eng. **3**, 466–477 (2019).
31. A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," Optica **4**, 1117–1125 (2017).
32. S. Li, M. Deng, J. Lee, A. Sinha, and G. Barbastathis, "Imaging through glass diffusers using densely connected convolutional networks," Optica **5**, 803 (2018).
33. J. Wu, L. Cao, L. Cao, G. Barbastathis, and G. Barbastathis, "DNN-FZA camera: a deep learning approach toward broadband FZA lensless imaging," Opt. Lett. **46**, 130–133 (2021).
34. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, 2016).

35. S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vucković, and A. W. Rodriguez, "Inverse design in nanophotonics," Nat. Photonics **12**, 659–670 (2018).
36. D. D. El-Mosalmy, M. F. O. Hameed, N. F. F. Areed, and S. S. A. Obayya, "Novel neural network based optimization approach for photonic devices," Opt. Quantum Electron. **46**, 439–453 (2014).
37. J. Peurifoy, Y. Shen, L. Jing, Y. Yang, F. Cano-Renteria, B. G. DeLacy, J. D. Joannopoulos, M. Tegmark, and M. Soljačić, "Nanophotonic particle simulation and inverse design using artificial neural networks," Sci. Adv. **4**, eaar4206 (2018).
38. D. Liu, Y. Tan, E. Khoram, and Z. Yu, "Training deep neural networks for the inverse design of nanophotonic structures," ACS Photonics **5**, 1365–1369 (2018).
39. 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**, 1004–1008 (2018).
40. J. Li, D. Mengu, N. T. Yardimci, Y. Luo, X. Li, M. Veli, Y. Rivenson, M. Jarrahi, and A. Ozcan, "Spectrally encoded single-pixel machine vision using diffractive networks," Sci. Adv. **7**, eabd7690 (2021).
41. Y. Luo, D. Mengu, N. T. Yardimci, Y. Rivenson, M. Veli, M. Jarrahi, and A. Ozcan, "Design of task-specific optical systems using broadband diffractive neural networks," Light: Sci. Appl. **8**, 112 (2019).
42. M. Veli, D. Mengu, N. T. Yardimci, Y. Luo, J. Li, Y. Rivenson, M. Jarrahi, and A. Ozcan, "Terahertz pulse shaping using diffractive surfaces," arXiv:2006.16599 [physics] (2020).
43. V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," ACM Trans. Graph. **37**, 1–13 (2018).
44. C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020), pp. 1372–1382.
45. L. Hu, S. Hu, W. Gong, and K. Si, "Learning-based Shack-Hartmann wavefront sensor for high-order aberration detection," Opt. Express **27**, 33504–33517 (2019).
46. I. Vishniakou and J. D. Seelig, "Wavefront correction for adaptive optics with reflected light and deep neural networks," Opt. Express **28**, 15459–15471 (2020).
47. W. S. Mcculloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Math. Biophys. **5**, 115–133 (1943).
48. F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," Psychol. Rev. **65**, 386 (1958).
49. G. Cybenkot, "Approximation by superpositions of a sigmoidal function," Math. Control Signal Systems **2**, 303–314 (1989).
50. K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Netw. **2**, 359–366 (1989).
51. D. E. Rumelhart, G. E. Hintont, and R. J. Williams, "Learning representations by back-propagating errors," Nature **323**, 533–536 (1986).
52. M. Leshno, V. Ya, A. P. Lin, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," Neural Netw. **6**, 861–867 (1993).
53. S. Sonoda and N. Murata, "Neural network with unbounded activation functions is universal approximator," Appl. Comput. Harmon. Anal. **43**, 233–268 (2017).
54. F. Voigtlaender, "The universal approximation theorem for complex-valued neural networks," arXiv:2012.03351v1.
55. N. L. Roux and Y. Bengio, "Deep belief networks are compact universal approximators," Neural Comput. **22**, 2192–2207 (2010).

56. G. F. Montúfar, "Universal approximation depth and errors of narrow Belief networks with discrete units," arXiv:1303.7461 [cs, math, stat] (2014).
57. D. P. Kingma and J. Ba, "Adam: amethod for stochastic optimization," arXiv:1412.6980 [cs] (2014).
58. L. Bottou, "Stochastic gradient learning in neural networks," 12 (n.d.).
59. Y. LeCun, L. Bottou, Y. Bengio, and P. Ha, "Gradient-based learning applied to document recognition," Proc. IEEE **86**, 2278–2374 (1998).
60. J. W. Goodman, *Introduction to Fourier Optics* (Roberts and Company Publishers, 2005).
61. H. M. Ozaktas, D. Mendlovic, M. A. Kutay, and Z. Zalevsky, *The Fractional Fourier Transform: With Applications in Optics and Signal Processing* (Wiley, 2001).
62. E. O'Neill, "Spatial filtering in optics," IEEE Trans. Inf. Theory **2**, 56–65 (1956).
63. L. Cutrona, E. Leith, C. Palermo, and L. Porcello, "Optical data processing and filtering systems," IEEE Trans. Inf. Theory **6**, 386–400 (1960).
64. A. V. Lugt, "Signal detection by complex spatial filtering," IEEE Trans. Inf. Theory **10**, 139–145 (1964).
65. C. S. Weaver and J. W. Goodman, "A technique for optically convolving two functions," Appl. Opt. **5**, 1248 (1966).
66. P. Refregier, B. Javidi, and V. Laude, "Nonlinear joint-transform correlation: an optimal solution for adaptive image discrimination and input noise robustness," Opt. Lett. **19**, 405 (1994).
67. B. Javidi, "Comparison of the nonlinear joint transform correlator and the non-linearly transformed matched filter based correlator for noisy input scenes," Opt. Eng. **29**, 55703 (1990).
68. S. Liu, C. Guo, and J. T. Sheridan, "A review of optical image encryption techniques," Opt. Laser Technol. **57**, 327–342 (2014).
69. W. Chen, X. Chen, and C. J. R. Sheppard, "Optical image encryption based on diffractive imaging," Opt. Lett. **35**, 3817 (2010).
70. P. Refregier and B. Javidi, "Optical image encryption based on input plane and fourier plane random encoding," Opt. Lett. **20**, 767 (1995).
71. E. N. Leith, "The evolution of information optics," IEEE J. Sel. Top. Quantum Electron. **6**, 1297–1304 (2000).
72. E. N. Leith, "Optical processing techniques for simultaneous pulse compression and beam sharpening," IEEE Trans. Aerosp. Electron. Syst. **AES-4**, 879–885 (1968).
73. A. Kozma, E. N. Leith, and N. G. Massey, "Tilted-plane optical processor," Appl. Opt. **11**, 1766 (1972).
74. L. J. Cutrona, E. N. Leith, L. J. Porcello, and W. E. Vivian, "On the application of coherent optical processing techniques to synthetic-aperture radar," Proc. IEEE **54**, 1026–1032 (1966).
75. D. Casasent, "Coherent optical pattern recognition," Proc. IEEE **67**, 813–825 (1979).
76. U. Mahlab, H. J. Caulfield, and J. Shamir, "Genetic algorithm for optical pattern recognition," Opt. Lett. **16**, 648 (1991).
77. Y.-N. Hsu and H. H. Arsenault, "Optical pattern recognition using circular harmonic expansion," Appl. Opt. **21**, 4016 (1982).
78. B. Javidi, "Optical pattern recognition for validation and security verification," Opt. Eng. **33**, 170736 (1994).
79. P. Ambs, S. H. Lee, Q. Tian, and Y. Fainman, "Optical implementation of the hough transform by a matrix of holograms," Appl. Opt. **25**, 4039–4045 (1986).

80. J. W. Goodman, "Linear space-variant optical data processing," in *Optical Information Processing*, S. H. Lee, ed., Topics in Applied Physics (Springer Berlin Heidelberg, 1981), Vol. 48, pp. 235–260.
81. W. Schneider and W. Fink, "Incoherent optical matrix multiplication," Opt. Acta **22**, 879–889 (1975).
82. J. W. Goodman and L. M. Woody, "Method for performing complex-valued linear operations on complex-valued data using incoherent light," Appl. Opt. **16**, 2611–2612 (1977).
83. J. Armitage and A. Lohmann, "Character recognition by incoherent spatial filtering," Appl. Opt. **4**, 461–467 (1965).
84. J. W. Goodman, A. R. Dias, and L. M. Woody, "Fully parallel, high-speed incoherent optical method for performing discrete fourier transforms," Opt. Lett. **2**, 1 (1978).
85. W. Rhodes and A. Sawchuk, "Incoherent optical processing," in *Optical Information Processing* (Springer, 1981), pp. 69–110.
86. G. Keryer, "On-board optical joint transform correlator for real-time road sign recognition," Opt. Eng. **34**, 135 (1995).
87. P. M. Birch, "Optical design of a miniature fourier transform lens system for a hybrid digital-optical correlator," Opt. Eng. **41**, 1650 (2002).
88. C. Mead, "Neuromorphic electronic systems," Proc. IEEE **78**, 8 (1990).
89. C. Mead, "Adaptive retina," in *Analog VLSI Implementation of Neural Systems* (Springer, 1989), pp. 239–246.
90. F. Blayo and P. Hurat, "A VLSI systolic array dedicated to Hopfield neural network," in *VLSI for Artificial Intelligence*, J. G. Delgado-Frias and W. R. Moore, eds., The Kluwer International Series in Engineering and Computer Science (Springer US, 1989), pp. 255–264.
91. G. Cauwenberghs, "An analog VLSI recurrent neural network learning a continuous-time trajectory," IEEE Trans. Neural Netw. **7**, 346–361 (1996).
92. M. A. C. Maher, S. P. Deweerth, M. A. Mahowald, and C. A. Mead, "Implementing neural architectures using analog VLSI circuits," IEEE Trans. Circuits Syst. **36**, 643–652 (1989).
93. Y. S. Abu-Mostafa and D. Psaltis, "Optical neural computers," Sci. Am. **256**, 88–95 (1987).
94. J. W. Goodman, F. J. Leonberger, S.-Y. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," Proc. IEEE **72**, 850–866 (1984).
95. B. Javidi, J. Li, and Q. Tang, "Optical implementation of neural networks for face recognition by the use of nonlinear joint transform correlators," Appl. Opt. **34**, 3950–3962 (1995).
96. K. Wagner and D. Psaltis, "Multilayer optical learning networks," Appl. Opt. **26**, 5061–5076 (1987).
97. D. Psaltis, D. Brady, and K. Wagner, "Adaptive optical networks using photorefractive crystals," Appl. Opt. **27**, 1752–1759 (1988).
98. N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical implementation of the hopfield model," Appl. Opt. **24**, 1469–1475 (1985).
99. H. Rajbenbach, Y. Fainman, and S. H. Lee, "Optical implementation of an iterative algorithm for matrix inversion," Appl. Opt. **26**, 1024 (1987).
100. H. J. Caulfield, W. T. Rhodes, M. J. Foster, and S. Horvitz, "Optical implementation of systolic array processing," Opt. Commun. **40**, 86–90 (1981).
101. D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, "Holography in artificial neural networks," in *Landmark Papers on Photorefractive Nonlinear Optics* (WORLD SCIENTIFIC, 1995), pp. 541–546.
102. H.-Y. S. Li, Y. Qiao, and D. Psaltis, "Optical network for real-time face recognition," Appl. Opt. **32**, 5026–5035 (1993).

103. G. E. Hinton, "Reducing the dimensionality of data with neural networks," Science **313**, 504–507 (2006).
104. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM **60**, 84–90 (2017).
105. S. E. Miller, "Integrated optics: an introduction," The Bell Syst. Tech. J. **48**, 2059–2069 (1969).
106. E. H. Turner, "High-frequency electro-optic coefficients of lithium niobate," Appl. Phys. Lett. **8**, 303 (1966).
107. G. T. Reed, W. R. Headley, and C. E. J. Png, "Silicon photonics – the early years," *Optoelectronic Integration on Silicon II*, San Jose, California, 2005.
108. R. Soref and J. Larenzo, "All-silicon active and passive guided-wave components for $\lambda = 1.3$ and 1.6 μm," IEEE J. Quantum Electron. **22**, 873–879 (1986).
109. D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," Proc. IEEE **88**, 728–749 (2000).
110. D. A. B. Miller, "Attojoule optoelectronics for low-energy information processing and communications," J. Lightwave Technol. **35**, 346–396 (2017).
111. P. Dong, Y.-K. Chen, G.-H. Duan, and D. T. Neilson, "Silicon photonic devices and integrated circuits," Nanophotonics **3**, 215–228 (2014).
112. Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: a review," Optica **5**, 1354 (2018).
113. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**, 102–114 (2021).
114. M. A. Nahmias, T. F. de Lima, A. N. Tait, H. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic multiply-accumulate operations for neural networks," IEEE J. Sel. Top. Quantum Electron. **26**, 1–18 (2020).
115. M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," Phys. Rev. Lett. **73**, 58–61 (1994).
116. J. Carolan, C. Harrold, C. Sparrow, E. Martín-López, N. J. Russell, J. W. Silverstone, P. J. Shadbolt, N. Matsuda, M. Oguma, M. Itoh, G. D. Marshall, M. G. Thompson, J. C. F. Matthews, T. Hashimoto, J. L. O'Brien, and A. Laing, "Universal linear optics," Science **349**, 711–716 (2015).
117. W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," Optica **3**, 1460–1465 (2016).
118. 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**, 441–446 (2017).
119. K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman, "Parallel reservoir computing using optical amplifiers," IEEE Trans. Neural Netw. **22**, 1469–1481 (2011).
120. M. J. Connelly, *Semiconductor Optical Amplifiers* (Springer Science & Business Media, 2007).
121. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**, 208 (2019).
122. A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**, 7430 (2017).
123. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, "11 TOPS photonic convolutional accelerator for optical neural networks," Nature **589**, 44–51 (2021).

124. F. Shokraneh, S. Geoffroy-gagnon, and O. Liboiron-Ladouceur, "The diamond mesh, a phase-error- and loss-tolerant field-programmable MZI-based optical processor for optical neural networks," Opt. Express **28**, 23495–23508 (2020).
125. M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," Opt. Express **27**, 14009–14029 (2019).
126. C. M. Wilkes, X. Qiang, J. Wang, R. Santagati, S. Paesani, X. Zhou, D. A. B. Miller, G. D. Marshall, M. G. Thompson, and J. L. O'Brien, "60 dB high-extinction auto-configured Mach–Zehnder interferometer," Opt. Lett. **41**, 5318 (2016).
127. D. A. B. Miller, "Self-configuring universal linear optical component [Invited]," Photonics Res. **1**, 1 (2013).
128. T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," Optica **5**, 864–871 (2018).
129. N. C. Harris, Y. Ma, J. Mower, T. Baehr-Jones, D. Englund, M. Hochberg, and C. Galland, "Efficient, compact and low loss thermo-optic phase shifter in silicon," Opt. Express **22**, 10487 (2014).
130. H. Jayatilleka, K. Murray, MÁ Guillén-Torres, M. Caverley, R. Hu, N. A. F. Jaeger, L. Chrostowski, and S. Shekhar, "Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters," Opt. Express **23**, 25084 (2015).
131. T. Komljenovic, M. Davenport, J. Hulme, A. Y. Liu, C. T. Santis, A. Spott, S. Srinivasan, E. J. Stanton, C. Zhang, and J. E. Bowers, "Heterogeneous silicon photonic integrated circuits," J. Lightwave Technol. **34**, 20–35 (2016).
132. M. He, M. Xu, Y. Ren, J. Jian, Z. Ruan, Y. Xu, S. Gao, S. Sun, X. Wen, L. Zhou, L. Liu, C. Guo, H. Chen, S. Yu, L. Liu, and X. Cai, "High-performance hybrid silicon and lithium niobate Mach–Zehnder modulators for 100 Gbit s−1 and beyond," Nat. Photonics **13**, 359–364 (2019).
133. V. Sorianello, M. Midrio, G. Contestabile, I. Asselberghs, J. Van Campenhout, C. Huyghebaert, I. Goykhman, A. K. Ott, A. C. Ferrari, and M. Romagnoli, "Graphene–silicon phase modulators with gigahertz bandwidth," Nat. Photonics **12**, 40–44 (2018).
134. C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "Integrated all-photonic non-volatile multi-level memory," Nat. Photonics **9**, 725–732 (2015).
135. B. Gholipour, P. Bastock, C. Craig, K. Khan, D. Hewak, and C. Soci, "Amorphous metal-sulphide microfibers enable photonic synapses for brain-like computing," Adv. Opt. Mater. **3**, 635–641 (2015).
136. A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: an integrated network for scalable photonic spike processing," J. Lightwave Technol. **32**, 4029–4041 (2014).
137. V. Bangari, B. A. Marquez, H. B. Miller, A. N. Tait, M. A. Nahmias, T. F. de Lima, H.-T. Peng, P. R. Prucnal, and B. J. Shastri, "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)," IEEE J. Sel. Top. Quantum Electron. **26**, 1–13 (2020).
138. B. Shi, N. Calabretta, and R. Stabile, "Deep neural network through an inp soa-based photonic integrated cross-connect," IEEE J. Sel. Top. Quantum Electron. **26**, 1–11 (2020).
139. X. Xu, M. Tan, B. Corcoran, J. Wu, T. G. Nguyen, A. Boes, S. T. Chu, B. E. Little, R. Morandotti, A. Mitchell, D. G. Hicks, and D. J. Moss, "Photonic perceptron based on a kerr microcomb for high-speed, scalable, optical neural networks," Laser Photonics Rev. **14**, 2000070 (2020).

140. M. A. Nahmias, H.-T. Peng, T. F. de Lima, C. Huang, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "A laser spiking neuron in a photonic integrated circuit," arXiv:2012.08516 [physics] (2020).
141. W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "HolyLight: a nanophotonic accelerator for deep learning in data centers," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)* (2019), pp. 1483–1488.
142. D. Dang, J. Dass, and R. Mahapatra, "ConvLight: a convolutional accelerator with memristor integrated photonic computing," in *2017 IEEE 24th International Conference on High Performance Computing (HiPC)* (2017), pp. 114–123.
143. D. Dang, A. Khansama, R. Mahapatra, and D. Sahoo, "BPhoton-CNN: an ultra-fast photonic backpropagation accelerator for deep learning," in *Proceedings of the 2020 on Great Lakes Symposium on VLSI, GLSVLSI '20* (Association for Computing Machinery, 2020), pp. 27–32.
144. A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. El-Ghazawi, "PCNNA: a photonic convolutional neural network accelerator," in *2018 31st IEEE International System-on-Chip Conference (SOCC)* (2018), pp. 169–173.
145. S. Kumar, R. S. Williams, and Z. Wang, "Third-order nanocircuit elements for neuromorphic engineering," Nature **585**, 518–523 (2020).
146. B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, "neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations," Proc. IEEE **102**, 699–716 (2014).
147. F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **34**, 1537–1557 (2015).
148. S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," Proc. IEEE **102**, 652–665 (2014).
149. 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**, 668–673 (2014).
150. T. F. de Lima, B. J. Shastri, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, "Progress in neuromorphic photonics," Nanophotonics **6**, 577–599 (2017).
151. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, "Neuromorphic photonics, principles of," in *Encyclopedia of Complexity and Systems Science*, R. A. Meyers, ed. (Springer, 2018), pp. 1–37.
152. M. A. Nahmias, B. J. Shastri, A. N. Tait, T. F. de Lima, and P. R. Prucnal, "Neuromorphic photonics," Opt. Photon. News **29**(1), 34–41 (2018).
153. A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," The Journal of Physiology **117**, 500–544 (1952).
154. R. W. Keyes, "Optical logic-in the light of computer technology," Opt. Acta **32**, 525–535 (1985).
155. J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, "Superconducting optoelectronic circuits for neuromorphic computing," Phys. Rev. Appl. **7**, 034013 (2017).
156. A. N. McCaughan, V. B. Verma, S. M. Buckley, J. P. Allmaras, A. G. Kozorezov, A. N. Tait, S. W. Nam, and J. M. Shainline, "A superconducting thermal switch

with ultrahigh impedance for interfacing superconductors to semiconductors," Nat. Electron. **2**, 451–456 (2019).
157. M. A. Nahmias, A. N. Tait, L. Tolias, M. P. Chang, T. Ferreira de Lima, B. J. Shastri, and P. R. Prucnal, "An integrated analog O/E/O link for multi-channel laser neurons," Appl. Phys. Lett. **108**, 151106 (2016).
158. A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H. B. Miller, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Silicon photonic modulator neuron," Phys. Rev. Appl. **11**, 064043 (2019).
159. R. Amin, J. K. George, S. Sun, T. Ferreira de Lima, A. N. Tait, J. B. Khurgin, M. Miscuglio, B. J. Shastri, P. R. Prucnal, T. El-Ghazawi, and V. J. Sorger, "ITO-based electro-absorption modulator for photonic neural activation function," APL Mater. **7**, 081112 (2019).
160. J. K. George, A. Mehrabian, R. Amin, J. Meng, T. F. de Lima, A. N. Tait, B. J. Shastri, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, "Neuromorphic photonics with electro-absorption modulators," Opt. Express **27**, 5181–5191 (2019).
161. I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," IEEE J. Sel. Top. Quantum Electron. **26**, 1–12 (2020).
162. M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, "All-optical nonlinear activation function for photonic neural networks [Invited]," Opt. Mater. Express **8**, 3851–3863 (2018).
163. M. T. Hill, E. E. E. Frietman, H. de Waardt, G. Khoe, and H. J. S. Dorren, "All fiber-optic neural network using coupled SOA based ring lasers," IEEE Trans. Neural Netw. **13**, 1504–1513 (2002).
164. D. Rosenbluth, K. Kravtsov, M. P. Fok, and P. R. Prucnal, "A high performance photonic pulse processing device," Opt. Express **17**, 22767 (2009).
165. K. S. Kravtsov, M. P. Fok, P. R. Prucnal, and D. Rosenbluth, "Ultrafast all-optical implementation of a leaky integrate-and-fire neuron," Opt. Express **19**, 2133 (2011).
166. A. Sebastian, M. Le Gallo, G. W. Burr, S. Kim, M. BrightSky, and E. Eleftheriou, "Tutorial: brain-inspired computing using phase-change memory devices," J. Appl. Phys. **124**, 111101 (2018).
167. C. Huang, T. F. de Lima, A. Jha, S. Abbaslou, B. J. Shastri, and P. R. Prucnal, "Giant enhancement in signal contrast using integrated all-optical nonlinear thresholder," in *Optical Fiber Communication Conference (OFC) 2019* (OSA, 2019), p. M3E.2.
168. K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, and M. Notomi, "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," Nat. Photonics **13**, 454–459 (2019).
169. C. Huang and S. Fujisawa, T. F. A. N. de Lima, E. Tait, Y. Blow, S. Tian, A. Bilodeau, F. Jha, H. G. Yaman, H.-T. Batshon, B. J. Peng, Y. Shastri, T. Inada, Paul. R. Wang, and Prucnal, "Demonstration of photonic neural network for fiber nonlinearity compensation in long-haul transmission systems," in *Optical Fiber Communication Conference Postdeadline Papers 2020* (Optical Society of America, 2020), p. Th4C.6.
170. S. Ostojic, "Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons," Nat. Neurosci. **17**, 594–600 (2014).
171. A. Kumar, S. Rotter, and A. Aertsen, "Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding," Nat. Rev. Neurosci. **11**, 615–627 (2010).
172. M. Diesmann, M.-O. Gewaltig, and A. Aertsen, "Stable propagation of synchronous spiking in cortical neural networks," Nature **402**, 529–533 (1999).

173. P. R. Prucnal, B. J. Shastri, T. F. de Lima, M. A. Nahmias, and A. N. Tait, "Recent progress in semiconductor excitable lasers for photonic spike processing," Adv. Opt. Photonics **8**, 228 (2016).
174. A. Hurtado and J. Javaloyes, "Controllable spiking patterns in long-wavelength vertical cavity surface emitting lasers for neuromorphic photonics systems," Appl. Phys. Lett. **107**, 241103 (2015).
175. S. Xiang, Z. Ren, Y. Zhang, Z. Song, and Y. Hao, "All-optical neuromorphic XOR operation with inhibitory dynamics of a single photonic spiking neuron based on a VCSEL-SA," Opt. Lett. **45**, 1104 (2020).
176. J. Robertson, M. Hejda, J. Bueno, and A. Hurtado, "Ultrafast optical integration and pattern classification for neuromorphic photonics based on spiking VCSEL neurons," Sci. Rep. **10**, 6098 (2020).
177. M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," IEEE J. Sel. Top. Quantum Electron. **19**, 1–12 (2013).
178. S. Barbay, R. Kuszelewicz, and A. M. Yacomotti, "Excitability in a semiconductor laser with saturable absorber," Opt. Lett. **36**, 4476 (2011).
179. V. N. Chizhevsky, V. A. Kulchitsky, and S. Y. Kilin, "Artificial spiking neuron based on a single-photon avalanche diode and a microcavity laser," Appl. Phys. Lett. **119**, 041107 (2021).
180. W. Coomans, L. Gelens, S. Beri, J. Danckaert, and G. Van der Sande, "Solitary and coupled semiconductor ring lasers as optical spiking neurons," Phys. Rev. E **84**, 036209 (2011).
181. M. Brunstein, A. M. Yacomotti, I. Sagnes, F. Raineri, L. Bigot, and A. Levenson, "Excitability and self-pulsing in a photonic crystal nanocavity," Phys. Rev. A **85**, 031803 (2012).
182. S. Wieczorek, B. Krauskopf, and D. Lenstra, "Multipulse excitability in a semiconductor laser with optical injection," Phys. Rev. Lett. **88**, 063901 (2002).
183. K. Alexander, T. V. Vaerenbergh, M. Fiers, P. Mechet, J. Dambre, and P. Bienstman, "Excitability in optically injected microdisk lasers with phase controlled excitatory and inhibitory response," Opt. Express **21**, 26182–26191 (2013).
184. F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay, "Relative refractory period in an excitable semiconductor laser," Phys. Rev. Lett. **112**, 183902 (2014).
185. F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay, "Temporal summation in a neuromimetic micropillar laser," Opt. Lett. **40**, 5690 (2015).
186. B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene excitable laser," Sci. Rep. **6**, 19126 (2016).
187. P. Y. Ma, B. J. Shastri, T. F. de Lima, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, "All-optical digital-to-spike conversion using a graphene excitable laser," Opt. Express **25**, 33504 (2017).
188. M. Turconi, B. Garbin, M. Feyereisen, M. Giudici, and S. Barland, "Control of excitable pulses in an injection-locked semiconductor laser," Phys. Rev. E **88**, 022923 (2013).
189. T. Sorrentino, C. Quintero-Quiroz, A. Aragoneses, M. C. Torrent, and C. Masoller, "Effects of periodic forcing on the temporally correlated spikes of a semiconductor laser with feedback," Opt. Express **23**, 5571 (2015).
190. B. Romeira, J. Javaloyes, C. N. Ironside, J. M. L. Figueiredo, S. Balle, and O. Piro, "Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors," Opt. Express **21**, 20931 (2013).

191. S. Xu, J. Wang, R. Wang, J. Chen, and W. Zou, "High-accuracy optical convolution unit architecture for convolutional neural networks by cascaded acousto-optical modulator arrays," Opt. Express **27**, 19778–19787 (2019).
192. A. Borst and F. E. Theunissen, "Information theory and neural coding," Nat. Neurosci. **2**, 947–957 (1999).
193. R. Sarpeshkar, "Analog versus digital: extrapolating from electronics to neurobiology," Neural Comput. **10**, 1601–1638 (1998).
194. S. Thorpe, A. Delorme, and R. Van Rullen, "Spike-based strategies for rapid processing," Neural Netw. **14**, 715–725 (2001).
195. J. H. Lee, T. Delbruck, and M. Pfeiffer, "Training deep spiking neural networks using backpropagation," Front. Neurosci. **10**, 508 (2016).
196. N. G. Pavlidis, O. K. Tasoulis, V. P. Plagianakos, G. Nikiforidis, and M. N. Vrahatis, "Spiking neural network training using evolutionary algorithms," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005* (2005), 4 pp. 2190–2194.
197. J. P. Dominguez-Morales, Q. Liu, R. James, D. Gutierrez-Galan, A. Jimenez-Fernandez, S. Davidson, and S. Furber, "Deep spiking neural network model for time-variant signals classification: a real-time speech recognition approach," in *2018 International Joint Conference on Neural Networks (IJCNN)* (2018), pp. 1–8.
198. J. J. Wade, L. J. McDaid, J. A. Santos, and H. M. Sayers, "SWAT: a spiking neural network training algorithm for classification problems," IEEE Trans. Neural Netw. **21**, 1817–1830 (2010).
199. P. O'Connor, E. Gavves, and M. Welling, "Training a spiking neural network with equilibrium propagation," in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019), pp. 1516–1523.
200. J. Wu, E. Yılmaz, M. Zhang, H. Li, and K. C. Tan, "Deep spiking neural networks for large vocabulary automatic speech recognition," Front. Neurosci. **14**, 199 (2020).
201. S. Singh, D. Gupta, R. S. Anand, and V. Kumar, "Nonsubsampled shearlet based CT and MR medical image fusion using biologically inspired spiking neural network," Biomedical Signal Processing and Control **18**, 91–101 (2015).
202. B. Meftah, O. Lezoray, and A. Benyettou, "Segmentation and edge detection based on spiking neural network model," Neural Process Lett **32**, 131–146 (2010).
203. L. Cheng, Y. Liu, Z.-G. Hou, M. Tan, D. Du, and M. Fei, "A rapid spiking neural network approach with an application on hand gesture recognition," IEEE Trans. Cogn. Dev. Syst. **13**, 151–161 (2021).
204. L. Deng, Y. Wu, X. Hu, L. Liang, Y. Ding, G. Li, G. Zhao, P. Li, and Y. Xie, "Rethinking the performance comparison between SNNS and ANNS," Neural Netw. **121**, 294–307 (2020).
205. H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu, "An optical neural chip for implementing complex-valued neural network," Nat. Commun. **12**, 457 (2021).
206. P. Virtue, S. X. Yu, and M. Lustig, "Better than real: complex-valued neural nets for MRI fingerprinting," in *2017 IEEE International Conference on Image Processing (ICIP)* (2017), pp. 3953–3957.
207. C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," arXiv:1705.09792v4.
208. K. D. G. Maduranga, K. E. Helfrich, and Q. Ye, "Complex unitary recurrent neural networks using scaled cayley transform," AAAI **33**, 4528–4535 (2019).

209. D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt, "An experimental unification of reservoir computing methods," Neural Netw. **20**, 391–403 (2007).
210. H. Jaeger and H. Haas, "Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication," Science **304**, 78–80 (2004).
211. W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: a new framework for neural computation based on perturbations," Neural Comput. **14**, 2531–2560 (2002).
212. W. Banzhaf, ed., *Advances in Artificial Life: 7th European Conference, ECAL 2003* ; 2801, Dortmund, Germany, 14-17 September 2003 (Springer-Verlag, 2003).
213. H. Hauser, A. J. Ijspeert, R. M. Füchslin, R. Pfeifer, and W. Maass, "Towards a theoretical foundation for morphological computation with compliant bodies," Biol Cybern **105**, 355–370 (2011).
214. K. Nakajima, T. Li, H. Hauser, and R. Pfeifer, "Exploiting short-term memory in soft body dynamics as a computational resource," J. R. Soc. Interface. **11**, 20140437 (2014).
215. A. Lugnan, A. Katumba, F. Laporte, M. Freiberger, S. Sackesyn, C. Ma, E. Gooskens, J. Dambre, and P. Bienstman, "Photonic neuromorphic information processing and reservoir computing," APL Photonics **5**, 020901 (2020).
216. G. V. der Sande, D. Brunner, and M. C. Soriano, "Advances in photonic reservoir computing," Nanophotonics **6**, 561–576 (2017).
217. Y. K. Chembo, "Machine learning based on reservoir computing with time-delayed optoelectronic and photonic systems," Chaos **30**, 013111 (2020).
218. G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, "Recent advances in physical reservoir computing: a review," Neural Netw. **115**, 100–123 (2019).
219. M. Cucchi, C. Gruener, L. Petrauskas, P. Steiner, H. Tseng, A. Fischer, B. Penkovsky, C. Matthus, P. Birkholz, H. Kleemann, and K. Leo, "Reservoir computing with biocompatible organic electrochemical networks for brain-inspired biosignal classification," Sci. Adv. **7**, eabh0693 (2021).
220. Y. Kawai, J. Park, and M. Asada, "A small-world topology enhances the echo state property and signal propagation in reservoir computing," Neural Netw. **112**, 15–23 (2019).
221. A. Rodan and P. Tino, "Minimum complexity echo state network," IEEE Trans. Neural Netw. **22**, 131–144 (2011).
222. H. Jaeger, "A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach," GMD Report 159, German National Research Center for Information Technology, 2002.
223. I. B. Yildiz, H. Jaeger, and S. J. Kiebel, "Re-visiting the echo state property," Neural Netw. **35**, 1–9 (2012).
224. D. Brunner and I. Fischer, "Reconfigurable semiconductor laser networks based on diffractive coupling," Opt. Lett. **40**, 3854 (2015).
225. J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, "Reinforcement learning in a large-scale photonic recurrent neural network," Optica **5**, 756–760 (2018).
226. K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," Nat. Commun. **5**, 3541 (2014).
227. S. Sackesyn, C. Ma, J. Dambre, and P. Bienstman, "An enhanced architecture for silicon photonic reservoir computing," *Cognitive Computing 2018*, Hannover, Germany, 18 December 2018.

228. A. Katumba, J. Heyvaert, B. Schneider, S. Uvin, J. Dambre, and P. Bienstman, "Low-loss photonic reservoir computing with multimode photonic integrated circuits," Sci. Rep. **8**, 2653 (2018).
229. F. Laporte, A. Katumba, J. Dambre, and P. Bienstman, "Numerical demonstration of neuromorphic computing with photonic crystal cavities," Opt. Express **26**, 7955 (2018).
230. C. Mesaritakis, V. Papataxiarhis, and D. Syvridis, "Micro ring resonators as building blocks for an all-optical high-speed reservoir-computing bit-pattern-recognition system," J. Opt. Soc. Am. B **30**, 3048 (2013).
231. F. D.-L. Coarer, M. Sciamanna, A. Katumba, M. Freiberger, J. Dambre, P. Bienstman, and D. Rontani, "All-optical reservoir computing on a photonic chip using silicon-based ring resonators," IEEE J. Sel. Top. Quantum Electron. **24**, 1–8 (2018).
232. L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, "Information processing using a single dynamical node as complex system," Nat. Commun. **2**, 468 (2011).
233. L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, "Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing," Opt. Express **20**, 3241–3249 (2012).
234. Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, "Optoelectronic reservoir computing," Sci. Rep. **2**, 287 (2012).
235. F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, "Fully analogue photonic reservoir computer," Sci. Rep. **6**, 22381 (2016).
236. L. Larger, A. Baylón-Fuentes, R. Martinenghi, V. S. Udaltsov, Y. K. Chembo, and M. Jacquot, "High-speed photonic reservoir computing using a time-delay-based architecture: million words per second classification," Phys. Rev. X **7**, 011015 (2017).
237. R. Martinenghi, S. Rybalko, M. Jacquot, Y. K. Chembo, and L. Larger, "Photonic nonlinear transient computing with multiple-delay wavelength dynamics," Phys. Rev. Lett. **108**, 244101 (2012).
238. M. C. Soriano, S. Ortín, D. Brunner, L. Larger, C. R. Mirasso, I. Fischer, and L. Pesquera, "Optoelectronic reservoir computing: tackling noise-induced performance degradation," Opt. Express **21**, 12–20 (2013).
239. S. Ortín, M. C. Soriano, L. Pesquera, D. Brunner, D. San-Martín, I. Fischer, C. R. Mirasso, and J. M. Gutiérrez, "A unified framework for reservoir computing and extreme learning machines based on a single time-delayed neuron," Sci. Rep. **5**, 14945 (2015).
240. F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, "Virtualization of a photonic reservoir computer," J. Lightwave Technol. **34**, 2085–2091 (2016).
241. Y. Chen, L. Yi, J. Ke, Z. Yang, Y. Yang, L. Huang, Q. Zhuge, and W. Hu, "Reservoir computing system with double optoelectronic feedback loops," Opt. Express **27**, 27431 (2019).
242. P. Antonik, M. Haelterman, and S. Massar, "Brain-inspired photonic signal processor for generating periodic patterns and emulating chaotic systems," Phys. Rev. Appl. **7**, 054014 (2017).
243. A. Argyris, J. Bueno, and I. Fischer, "Photonic machine learning implementation for signal recovery in optical communications," Sci. Rep. **8**, 8487 (2018).
244. F. Duport, B. Schneider, A. Smerieri, M. Haelterman, and S. Massar, "All-optical reservoir computing," Opt. Express **20**, 22783–22795 (2012).
245. D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," Nat. Commun. **4**, 1364 (2013).

246. Q. Vinckier, F. Duport, A. Smerieri, K. Vandoorne, P. Bienstman, M. Haelterman, and S. Massar, "High-performance photonic reservoir computer based on a coherently driven passive cavity," Optica **2**, 438–446 (2015).
247. A. Dejonckheere, F. Duport, A. Smerieri, L. Fang, J.-L. Oudar, M. Haelterman, and S. Massar, "All-optical reservoir computer based on saturation of absorption," Opt. Express **22**, 10868 (2014).
248. K. Hicke, M. A. Escalona-Morán, D. Brunner, M. C. Soriano, I. Fischer, and C. R. Mirasso, "Information processing using transient dynamics of semiconductor lasers subject to delayed feedback," IEEE J. Sel. Top. Quantum Electron. **19**, 1501610 (2013).
249. Y. Kuriki, J. Nakayama, K. Takano, and A. Uchida, "Impact of input mask signals on delay-based photonic reservoir computing with semiconductor lasers," Opt. Express **26**, 5777 (2018).
250. R. M. Nguimdo, E. Lacot, O. Jacquin, O. Hugon, G. Van der Sande, and H. G. de Chatellus, "Prediction performance of reservoir computing systems based on a diode-pumped erbium-doped microchip laser subject to optical feedback," Opt. Lett. **42**, 375 (2017).
251. K. Takano, C. Sugano, M. Inubushi, K. Yoshimura, S. Sunada, K. Kanno, and A. Uchida, "Compact reservoir computing with a photonic integrated circuit," Opt. Express **26**, 29424 (2018).
252. J. Vatin, D. Rontani, and M. Sciamanna, "Experimental reservoir computing using VCSEL polarization dynamics," Opt. Express **27**, 18579 (2019).
253. J. Bueno, D. Brunner, M. C. Soriano, and I. Fischer, "Conditions for reservoir computing performance using semiconductor lasers with delayed optical feedback," Opt. Express **25**, 2401 (2017).
254. R. M. Nguimdo and T. Erneux, "Enhanced performances of a photonic reservoir computer based on a single delayed quantum cascade laser," Opt. Lett. **44**, 49 (2019).
255. C. Mesaritakis and D. Syvridis, "Reservoir computing based on transverse modes in a single optical waveguide," Opt. Lett. **44**, 1218–1221 (2019).
256. U. Paudel, M. Luengo-Kovac, J. Pilawa, T. J. Shaw, and G. C. Valley, "Classification of time-domain waveforms using a speckle-based optical reservoir computer," Opt. Express **28**, 1225 (2020).
257. U. Teğin, M. Yıldırım, İ. Oğuz, C. Moser, and D. Psaltis, "Scalable optical learning operator," Nat. Comput. Sci. **1**, 542549 (2021).
258. F. Zangeneh-Nejad, D. L. Sounas, A. Alù, and R. Fleury, "Analogue computing with metamaterials," Nat. Rev. Mater. **6**, 207–225 (2021).
259. N. M. Estakhri, B. Edwards, and N. Engheta, "Inverse-designed metastructures that solve equations," Science **363**, 1333–1338 (2019).
260. O. K. Ersoy, *Diffraction, Fourier Optics, and Imaging* (Wiley-Interscience, 2007).
261. F. Shen and A. Wang, "Fast-Fourier-transform based numerical integration method for the Rayleigh-Sommerfeld diffraction formula," Appl. Opt. **45**, 1102 (2006).
262. D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," IEEE J. Sel. Top. Quantum Electron. **26**, 1–14 (2020).
263. O. Kulce, D. Mengu, Y. Rivenson, and A. Ozcan, "All-optical information-processing capacity of diffractive surfaces," Light: Sci. Appl. **10**, 25 (2021).
264. O. Kulce, D. Mengu, Y. Rivenson, and A. Ozcan, "All-optical synthesis of an arbitrary linear transformation using diffractive surfaces," Light: Sci. Appl. **10**, 196 (2021).
265. D. Mengu, Y. Zhao, N. T. Yardimci, Y. Rivenson, M. Jarrahi, and A. Ozcan, "Misalignment resilient diffractive optical networks," Nanophotonics **9**, 1 (2020).

266. J. Li, D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Class-specific differential detection in diffractive optical neural networks improves inference accuracy," Adv. Photonics **1**, 046001 (2019).
267. M. S. S. Rahman, J. Li, D. Mengu, Y. Rivenson, and A. Ozcan, "Ensemble learning of diffractive optical networks," Light: Sci. Appl. **10**, 14 (2021).
268. H. Dou, Y. Deng, T. Yan, H. Wu, X. Lin, and Q. Dai, "Residual $D^2$NN: training diffractive deep neural networks via learnable light shortcuts," Opt. Lett. **45**, 2688–2691 (2020).
269. T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," Nat. Photonics **15**, 367–373 (2021).
270. J. Shi, L. Zhou, T. Liu, C. Hu, K. Liu, J. Luo, H. Wang, C. Xie, and X. Zhang, "Multiple-view $D^2$ NNs array: realizing robust 3D object recognition," Opt. Lett. **46**, 3388 (2021).
271. T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, and Q. Dai, "Fourier-space diffractive deep neural network," Phys. Rev. Lett. **123**, 023901 (2019).
272. D. Mengu, M. Veli, Y. Rivenson, and A. Ozcan, "Classification and reconstruction of spatially overlapping phase images using diffractive optical networks," arXiv:2108.07977 [physics] (2021).
273. E. Goi, X. Chen, Q. Zhang, B. P. Cumming, S. Schoenhardt, H. Luan, and M. Gu, "Nanoprinted high-neuron-density optical linear perceptrons performing near-infrared inference on a CMOS chip," Light: Sci. Appl. **10**, 40 (2021).
274. C. Qian, X. Lin, X. Lin, J. Xu, Y. Sun, E. Li, B. Zhang, and H. Chen, "Performing optical logic operations by a diffractive neural network," Light: Sci. Appl. **9**, 1–7 (2020).
275. Y. Luo, D. Mengu, and A. Ozcan, "Cascadable all-optical NAND gates using diffractive networks," Sci. Rep. **12**, 7121 (2022).
276. T. Zhou, L. Fang, T. Yan, J. Wu, Y. Li, J. Fan, H. Wu, X. Lin, and Q. Dai, "In situ optical backpropagation training of diffractive optical neural networks," Photonics Res. **8**, 940 (2020).
277. A. S. Backer, "Computational inverse design for cascaded systems of metasurface optics," Opt. Express **27**, 30308 (2019).
278. O. Kulce, D. Mengu, Y. Rivenson, and A. Ozcan, "All-optical synthesis of an arbitrary linear transformation using diffractive surfaces," arXiv:2108.09833 [physics] (2021).
279. N. U. Dinc, J. Lim, E. Kakkava, C. Moser, and D. Psaltis, "Computer generated optical volume elements by additive manufacturing," Nanophotonics **9**, 4173–4181 (2020).
280. Z. Huang, P. Wang, J. Liu, W. Xiong, Y. He, J. Xiao, H. Ye, Y. Li, S. Chen, and D. Fan, "All-optical signal processing of vortex beams with diffractive deep neural networks," Phys. Rev. Appl. **15**, 014037 (2021).
281. 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**, 1132–1137 (2019).
282. J. Shi, J. Shi, M. Chen, M. Chen, D. Wei, D. Wei, C. Hu, C. Hu, C. Hu, J. Luo, J. Luo, H. Wang, X. Zhang, X. Zhang, X. Zhang, C. Xie, and C. Xie, "Anti-noise diffractive neural network for constructing an intelligent imaging detector array," Opt. Express **28**, 37686–37699 (2020).
283. D. Mengu, Y. Rivenson, and A. Ozcan, "Scale-, shift-, and rotation-invariant diffractive optical networks," ACS Photonics **8**, 324–334 (2021).
284. T. W. Hughes, I. A. D. Williamson, M. Minkov, and S. Fan, "Wave physics as an analog recurrent neural network," Sci. Adv. **5**, eaay6946 (2019).

285. E. Khoram, A. Chen, D. Liu, L. Ying, Q. Wang, M. Yuan, and Z. Yu, "Nanophotonic media for artificial neural inference," Photonics Res. **7**, 823 (2019).
286. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," Phys. Rev. X **9**, 021032 (2019).
287. P. Antonik, N. Marsal, and D. Rontani, "Large-scale spatiotemporal photonic reservoir computer for image classification," IEEE J. Sel. Top. Quantum Electron. **26**, 1–12 (2020).
288. J. Dong, M. Rafayelyan, F. Krzakala, and S. Gigan, "Optical reservoir computing using multiple light scattering for chaotic systems prediction," IEEE J. Sel. Top. Quantum Electron. **26**, 1–12 (2020).
289. J. Pauwels, G. Van der Sande, A. Bouwens, M. Haelterman, and S. Massar, "Towards high-performance spatially parallel optical reservoir computing," in *Neuro-Inspired Photonic Computing*, M. Sciamanna and P. Bienstman, eds. (SPIE, 2018), p. 3.
290. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Sci. Rep. **8**, 12324 (2018).
291. N. T. Yardimci and M. Jarrahi, "High sensitivity terahertz detection through large-area plasmonic nano-antenna arrays," Sci. Rep. **7**, 1–8 (2017).
292. G. Côté, G. Côté, J.-F. Lalonde, and S. Thibault, "Extrapolating from lens design databases using deep learning," Opt. Express **27**, 28279–28292 (2019).
293. G. Côté, J.-F. Lalonde, and S. Thibault, "Introducing a dynamic deep neural network to infer lens design starting points," in *Current Developments in Lens Design and Optical Engineering XX* (International Society for Optics and Photonics, 2019), Vol. 11104, p. 1110403.
294. G. Côté, G. Côté, J.-F. Lalonde, and S. Thibault, "Deep learning-enabled framework for automatic lens design starting point generation," Opt. Express **29**, 3841–3854 (2021).
295. A. F. Koenderink, A. Alù, and A. Polman, "Nanophotonics: shrinking light-based technology," Science **348**, 516–521 (2015).
296. L. Bianchi, M. Dorigo, L. M. Gambardella, and W. J. Gutjahr, "A survey on metaheuristics for stochastic combinatorial optimization," Nat. Comput. **8**, 239–287 (2009).
297. C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch, "Adjoint shape optimization applied to electromagnetic design," Opt. Express **21**, 21693–21701 (2013).
298. T.-S. Horng, C.-C. Wang, and N. G. Alexopoulos, "Microstrip circuit design using neural networks," in *1993 IEEE MTT-S International Microwave Symposium Digest (1993)*, pp. 413–416 vol. 1.
299. M. Vai and S. Prasad, "Automatic impedance matching with a neural network," IEEE Microw. Guid. Wave Lett. **3**, 353–354 (1993).
300. P. Burrascano, S. Fiori, and M. Mongiardo, "A review of artificial neural networks applications in microwave computer-aided design (invited article)," Int. J. RF and Microwave Comp. Aid. Eng. **9**, 158–174 (1999).
301. R. Unni, K. Yao, and Y. Zheng, "Deep convolutional mixture density network for inverse design of layered photonic structures," ACS Photonics **7**, 2703–2712 (2020).
302. S. So and J. Rho, "Designing nanophotonic structures using conditional deep convolutional generative adversarial networks," Nanophotonics **8**, 1255–1261 (2019).

303. W. Ma, F. Cheng, and Y. Liu, "Deep-learning-enabled on-demand design of chiral metamaterials," ACS Nano **12**, 6326–6334 (2018).
304. I. Malkiel, M. Mrejen, A. Nagler, U. Arieli, L. Wolf, and H. Suchowski, "Plasmonic nanostructure design and characterization via deep learning," Light: Sci. Appl. **7**, 60 (2018).
305. Z. Liu, D. Zhu, S. P. Rodrigues, K.-T. Lee, and W. Cai, "Generative model for the inverse design of metasurfaces," Nano Lett. **18**, 6570–6576 (2018).
306. T. Qiu, X. Shi, J. Wang, Y. Li, S. Qu, Q. Cheng, T. Cui, and S. Sui, "Deep learning: a rapid and efficient route to automatic metasurface design," Adv. Sci. **6**, 1900128 (2019).
307. Y. Chen, J. Zhu, Y. Xie, N. Feng, and Q. H. Liu, "Smart inverse design of graphene-based photonic metamaterials by an adaptive artificial neural network," Nanoscale **11**, 9749–9755 (2019).
308. C. C. Nadell, B. Huang, J. M. Malof, and W. J. Padilla, "Deep learning for accelerated all-dielectric metasurface design," Opt. Express **27**, 27523–27535 (2019).
309. S. An, C. Fowler, B. Zheng, M. Y. Shalaginov, H. Tang, H. Li, L. Zhou, J. Ding, A. M. Agarwal, C. Rivero-Baleine, K. A. Richardson, T. Gu, J. Hu, and H. Zhang, "A deep learning approach for objective-driven all-dielectric metasurface design," ACS Photonics **6**, 3196–3207 (2019).
310. W. Ma, F. Cheng, Y. Xu, Q. Wen, and Y. Liu, "Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy," Adv. Mater. **31**, 1901111 (2019).
311. Z. Liu, L. Raju, D. Zhu, and W. Cai, "A hybrid strategy for the discovery and design of photonic structures," IEEE J. Emerg. Sel. Topics Circuits Syst. **10**, 126–135 (2020).
312. X. Shi, T. Qiu, J. Wang, X. Zhao, and S. Qu, "Metasurface inverse design using machine learning approaches," J. Phys. D: Appl. Phys. **53**, 275105 (2020).
313. S. So, J. Mun, and J. Rho, "Simultaneous inverse design of materials and structures via deep learning: demonstration of dipole resonance engineering using Core–Shell nanoparticles," ACS Appl. Mater. Interfaces **11**, 24264–24268 (2019).
314. R. Singh, A. Agarwal, and B. W. Anthony, "Design of optical meta-structures with applications to beam engineering using deep learning," Sci. Rep. **10**, 19923 (2020).
315. C. Qian, B. Zheng, Y. Shen, L. Jing, E. Li, L. Shen, and H. Chen, "Deep-learning-enabled self-adaptive microwave cloak without human intervention," Nat. Photonics **14**, 383–390 (2020).
316. M. V. Zhelyeznyakov, S. Brunton, and A. Majumdar, "Deep learning to accelerate Scatterer-to-Field mapping for inverse design of dielectric metasurfaces," ACS Photonics **8**, 481 (2021).
317. Q. Zhang, C. Liu, X. Wan, L. Zhang, S. Liu, Y. Yang, and T. J. Cui, "Machine-learning designs of anisotropic digital coding metasurfaces," Adv. Theory Simul. **2**, 1800132 (2019).
318. Z. Liu, D. Zhu, K.-T. Lee, A. S. Kim, L. Raju, and W. Cai, "Compounding meta-atoms into metamolecules with hybrid artificial intelligence techniques," Adv. Mater. **32**, 1904790 (2020).
319. S. An, B. Zheng, H. Tang, M. Y. Shalaginov, L. Zhou, H. Li, M. Kang, K. A. Richardson, T. Gu, J. Hu, C. Fowler, and H. Zhang, "Multifunctional metasurface design with a generative adversarial network," Adv. Opt. Mater. **9**, 2001433 (2021).

320. S. Inampudi and H. Mosallaei, "Neural network based design of metagratings," Appl. Phys. Lett. **112**, 241102 (2018).
321. J. Jiang, D. Sell, S. Hoyer, J. Hickey, J. Yang, and J. A. Fan, "Free-form diffractive metagrating design based on generative adversarial networks," ACS Nano **13**, 8872–8878 (2019).
322. I. Sajedian, H. Lee, and J. Rho, "Double-deep Q-learning to increase the efficiency of metasurface holograms," Sci. Rep. **9**, 10899 (2019).
323. H. Ren, W. Shao, Y. Li, F. Salim, and M. Gu, "Three-dimensional vectorial holography based on machine learning inverse design," Sci. Adv. **6**, eaaz4261 (2020).
324. M. S. Sakib Rahman and A. Ozcan, "Computer-free, all-optical reconstruction of holograms using diffractive networks," ACS Photonics **8**, 3375 (2021).
325. L. Gao, X. Li, D. Liu, L. Wang, and Z. Yu, "A bidirectional deep neural network for accurate silicon color design," Adv. Mater. **31**, 1905467 (2019).
326. O. Hemmatyar, S. Abdollahramezani, Y. Kiarashinejad, M. Zandehshahvar, and A. Adibi, "Full color generation with Fano-type resonant HfO2 nanopillars designed by a deep-learning approach," Nanoscale **11**, 21266–21274 (2019).
327. V. Kalt, A. K. González-Alcalde, S. Es-Saidi, R. Salas-Montiel, S. Blaize, and D. Macías, "Metamodeling of high-contrast-index gratings for color reproduction," J. Opt. Soc. Am. A **36**, 79–88 (2019).
328. I. Sajedian, T. Badloe, and J. Rho, "Optimization of color generation from dielectric nanostructures using reinforcement learning," Opt. Express **27**, 5874–5883 (2019).
329. T. Badloe, I. Kim, and J. Rho, "Biomimetic ultra-broadband perfect absorbers optimised with reinforcement learning," Phys. Chem. Chem. Phys. **22**, 2337–2342 (2020).
330. I. Sajedian, H. Lee, and J. Rho, "Design of high transmission color filters for solar cells directed by deep Q-learning," Sol. Energy **195**, 670–676 (2020).
331. Z. A. Kudyshev, A. V. Kildishev, V. M. Shalaev, and A. Boltasseva, "Machine-learning-assisted metasurface design for high-efficiency thermal emitter optimization," Appl. Phys. Rev. **7**, 021407 (2020).
332. A. Jiang, Y. Osamu, and L. Chen, "Multilayer optical thin film design with deep Q learning," Sci. Rep. **10**, 12780 (2020).
333. I. Sajedian, T. Badloe, H. Lee, and J. Rho, "Deep Q-network to produce polarization-independent perfect solar absorbers: a statistical report," Nano Convergence **7**, 26 (2020).
334. H. Wang, Z. Zheng, C. Ji, and L. J. Guo, "Automated multi-layer optical design via deep reinforcement learning," Mach. Learn.: Sci. Technol. **2**, 025013 (2021).
335. A. da Silva Ferreira, C. H. da Silva Santos, M. S. Gonçalves, and H. E. Hernández Figueroa, "Towards an integrated evolutionary strategy and artificial neural network computational tool for designing photonic coupler devices," Appl. Soft Comput. **65**, 1–11 (2018).
336. E. Bor, O. Alparslan, M. Turduev, Y. S. Hanay, H. Kurt, S. Arakawa, and M. Murata, "Integrated silicon photonic device design by attractor selection mechanism based on artificial neural networks: optical coupler and asymmetric light transmitter," Opt. Express **26**, 29032–29044 (2018).
337. M. H. Tahersima, K. Kojima, T. Koike-Akino, D. Jha, B. Wang, C. Lin, and K. Parsons, "Deep neural network inverse design of integrated photonic power splitters," Sci. Rep. **9**, 1368 (2019).
338. T. Zhang, J. Wang, Q. Liu, J. Zhou, J. Dai, X. Han, Y. Zhou, and K. Xu, "Efficient spectrum prediction and inverse design for plasmonic waveguide systems based on artificial neural networks," Photonics Res. **7**, 368–380 (2019).

339. A. M. Hammond and R. M. Camacho, "Designing integrated photonic devices using artificial neural networks," Opt. Express **27**, 29620–29638 (2019).
340. Z. Ballard, C. Brown, A. M. Madni, and A. Ozcan, "Machine learning and computation-enabled intelligent sensor design," Nat Mach Intell **3**, 556–565 (2021).
341. A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016), Vol. 29.
342. S. Nie, L. Gu, Y. Zheng, A. Lam, N. Ono, and I. Sato, "Deeply learned filter response functions for hyperspectral reconstruction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 4767–4776.
343. M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, "DeepBinaryMask: learning a binary mask for video compressive sensing," Digit. Signal Process. **96**, 102591 (2020).
344. Y. Peng, Q. Sun, X. Dun, G. Wetzstein, W. Heidrich, and F. Heide, "Learned large field-of-view imaging with thin-plate optics," ACM Trans. Graph. **38**, 1–14 (2019).
345. Y. Ba, A. Gilbert, F. Wang, J. Yang, R. Chen, Y. Wang, L. Yan, B. Shi, and A. Kadambi, "Deep shape from polarization," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds., Lecture Notes in Computer Science (Springer International Publishing, 2020), Vol. 12369, pp. 554–571.
346. S. Su, F. Heide, G. Wetzstein, and W. Heidrich, "Deep end-to-end time-of-flight imaging," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 6383–6392.
347. M. Tancik, G. Satat, and R. Raskar, "Flash photography for data-driven hidden scene recovery," arXiv:1810.11710 [cs] (2018).
348. A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," ACM Trans. Graph. **26**, 70-es (2007).
349. G. R. Arce, D. J. Brady, L. Carin, H. Arguello, and D. S. Kittle, "Compressive coded aperture spectral imaging: an introduction," IEEE Signal Process. Mag. **31**, 105–115 (2014).
350. A. Greengard, Y. Y. Schechner, and R. Piestun, "Depth from diffracted rotation," Opt. Lett. **31**, 181–183 (2006).
351. B. Huang, W. Wang, M. Bates, and X. Zhuang, "Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy," Science **319**, 810–813 (2008).
352. S. R. P. Pavani, M. A. Thompson, J. S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. E. Moerner, "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function," Proc. Natl. Acad. Sci. **106**, 2995–2999 (2009).
353. S. Elmalem, R. Giryes, and E. Marom, "Learned phase coded aperture for the benefit of depth of field extension," Opt. Express **26**, 15316–15331 (2018).
354. U. Akpinar, E. Sahin, and A. Gotchev, "Learning optimal phase-coded aperture for depth of field extension," in *2019 IEEE International Conference on Image Processing (ICIP)* (2019), pp. 4315–4319.
355. Q. Sun, J. Zhang, X. Dun, B. Ghanem, Y. Peng, and W. Heidrich, "End-to-end learned, optically coded super-resolution spad camera," ACM Trans. Graph. **39**, 1–14 (2020).
356. H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom, "Depth estimation from a single image using deep learned phase coded mask," IEEE Trans. Comput. Imaging **4**, 298–310 (2018).

357. Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "PhaseCam3D — learning phase masks for passive single view depth estimation," in *2019 IEEE International Conference on Computational Photography (ICCP)* (2019), pp. 1–12.
358. J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2019), pp. 10192–10201.
359. Q. Sun, E. Tseng, Q. Fu, W. Heidrich, and F. Heide, "Learning rank-1 diffractive optics for single-shot high dynamic range imaging," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020), pp. 1383–1393.
360. X. Dun, X. Dun, H. Ikoma, G. Wetzstein, Z. Wang, Z. Wang, X. Cheng, X. Cheng, X. Cheng, Y. Peng, and Y. Peng, "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," Optica **7**, 913–922 (2020).
361. S.-H. Baek, H. Ikoma, D. S. Jeon, Y. Li, W. Heidrich, G. Wetzstein, and M. H. Kim, "Single-shot hyperspectral-depth imaging with learned diffractive optics," in *International Conference on Computer Vision (ICCV) 2021*, 11-17 October 2021.
362. Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara, "Learning to capture light fields through a coded aperture camera," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds., Lecture Notes in Computer Science (Springer International Publishing, 2018), Vol. 11211, pp. 431–448.
363. U. Akpinar, E. Sahin, and A. Gotchev, "Phase-coded computational imaging for accommodation-invariant near-eye displays," in *2020 IEEE International Conference on Image Processing (ICIP)* (2020), pp. 3159–3163.
364. U. Akpinar, E. Sahin, and A. Gotchev, "Computational multifocal near-eye display with hybrid refractive-diffractive optics," in *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)* (2020), pp. 1–6.
365. Y. Shechtman, S. J. Sahl, A. S. Backer, and W. E. Moerner, "Optimal point spread function design for 3d imaging," Phys. Rev. Lett. **113**, 133902 (2014).
366. Y. Shechtman, L. E. Weiss, A. S. Backer, S. J. Sahl, and W. E. Moerner, "Precise three-dimensional scan-free multiple-particle tracking over large axial ranges with tetrapod point spread functions," Nano Lett. **15**, 4194–4199 (2015).
367. E. Nehme, B. Ferdman, L. E. Weiss, T. Naor, D. Freedman, T. Michaeli, and Y. Shechtman, "Learning optimal wavefront shaping for multi-channel imaging," IEEE Trans. Pattern Anal. Mach. Intell. **43**, 2179–2192 (2021).
368. E. Hershko, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Multicolor localization microscopy and point-spread-function engineering by deep learning," Opt. Express **27**, 6158 (2019).
369. S.-H. Baek, H. Ikoma, D. S. Jeon, Y. Li, W. Heidrich, G. Wetzstein, and M. H. Kim, "End-to-end hyperspectral-depth imaging with learned diffractive optics," (2020).
370. M. G. L. Gustafsson, "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy," J. Microsc. **198**, 82–87 (2000).
371. W. Luo, A. Greenbaum, Y. Zhang, and A. Ozcan, "Synthetic aperture-based on-chip microscopy," Light: Sci. Appl. **4**, e261 (2015).
372. L. Tian, X. Li, K. Ramchandran, and L. Waller, "Multiplexed coded illumination for fourier ptychography with an LED array microscope," Biomed. Opt. Express **5**, 2376–2389 (2014).
373. Y. F. Cheng, M. Strachan, Z. Weiss, M. Deb, D. Carone, and V. Ganapati, "Illumination pattern design with deep learning for single-shot fourier ptychographic microscopy," Opt. Express **27**, 644–656 (2019).

374. A. Robey and V. Ganapati, "Optimal physical preprocessing for example-based super-resolution," Opt. Express **26**, 31333 (2018).
375. M. Kellman, E. Bostan, M. Chen, and L. Waller, "Data-driven design for fourier ptychographic microscopy," in *2019 IEEE International Conference on Computational Photography (ICCP) (2019)*, pp. 1–8.
376. M. R. Kellman, E. Bostan, N. A. Repina, and L. Waller, "Physics-based learned design: optimized coded-illumination for quantitative phase imaging," IEEE Trans. Comput. Imaging **5**, 344–353 (2019).
377. H.-A. Joung, Z. S. Ballard, J. Wu, D. K. Tseng, H. Teshome, L. Zhang, E. J. Horn, P. M. Arnaboldi, R. J. Dattwyler, O. B. Garner, D. Di Carlo, and A. Ozcan, "Point-of-care serodiagnostic test for early-stage lyme disease using a multiplexed paper-based immunoassay and machine learning," ACS Nano **14**, 229–240 (2020).
378. C. Brown, A. Goncharov, Z. S. Ballard, M. Fordham, A. Clemens, Y. Qiu, Y. Rivenson, and A. Ozcan, "Neural network-based on-chip spectroscopy using a scalable plasmonic encoder," ACS Nano **15**, 6305–6315 (2021).
379. Y. Luo, H.-A. Joung, S. Esparza, J. Rao, O. Garner, and A. Ozcan, "Quantitative particle agglutination assay for point-of-care testing using mobile holographic imaging and deep learning," Lab Chip **21**, 3550–3558 (2021).
380. B. Diederich, R. Wartmann, H. Schadwinkel, and R. Heintzmann, "Using machine-learning to optimize phase contrast in a low-cost cellphone microscope," PLoS One **13**, e0192937 (2018).
381. H. Pinkard, H. Baghdassarian, A. Mujal, E. Roberts, K. H. Hu, D. H. Friedman, I. Malenica, T. Shagam, A. Fries, K. Corbin, M. F. Krummel, and L. Waller, "Learned adaptive multiphoton illumination microscopy for large-scale immune response imaging," Nat. Commun. **12**, 1916 (2021).
382. A. Turpin, I. Vishniakou, and J. d Seelig, "Light scattering control in transmission and reflection with neural networks," Opt. Express **26**, 30911–30929 (2018).
383. J. R. P. Angel, P. Wizinowich, M. Lloyd-Hart, and D. Sandler, "Adaptive optics for array telescopes using neural-network techniques," Nature **348**, 221–224 (1990).
384. S. W. Paine and J. R. Fienup, "Machine learning for improved image-based wavefront sensing," Opt. Lett. **43**, 1235–1238 (2018).
385. Y. Nishizaki, M. Valdivia, R. Horisaki, K. Kitaguchi, M. Saito, J. Tanida, and E. Vera, "Deep learning wavefront sensing," Opt. Express **27**, 240–251 (2019).
386. Y. Luo, Y. Zhao, J. Li, E. etinta, Y. Rivenson, M. Jarrahi, and A. Ozcan, "Computational imaging without a computer: seeing through random diffusers at the speed of light," eLight **2**, 4 (2022).
387. D. Mengu and A. Ozcan, "Diffractive all-optical computing for quantitative phase imaging," arXiv:2201.08964 [physics] (2022).

**Deniz Mengu** received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, the M.Sc. degree in electrical electronic engineering from Koç University, Istanbul, Turkey. He is currently working toward the Ph.D. degree with the Electrical and Computer Engineering Department, University of California, Los Angeles, California, USA. His research focuses on computational imaging and sensing platforms, machine learning, and optics.

**Md Sadman Sakib Rahman** is a fourth year Graduate Student at the Ozcan Research Group in UCLA, Los Angeles, California, USA. He received his B.Sc. in 2014, and M.Sc. in 2017, both in electrical and electronic engineering from Bangladesh University of Engineering and Technology, Dhaka. His research interest lies broadly in photonics and currently in the use of deep learning for design of optical computing hardware.

**Yi Luo** received the B.S. degree in measurement, control technology, and instrumentation from Tsinghua University, Beijing, China, in 2016. He is currently working toward a Ph.D. degree with the Bioengineering Department, University of California, Los Angeles, California, USA. His work focuses on the development of computational imaging and sensing platforms.

**Jingxi Li** received his B.Sc. degree in opto-electronic information science and engineering from Zhejiang University, Hangzhou, Zhejiang, China, in 2018. Currently, he is working toward his Ph.D. with the Electrical and Computer Engineering Department at the University of California, Los Angeles, California, USA. His work focuses on computational optical imaging and sensing platforms.

**Onur Kulce** received his B.Sc., M.Sc., and Ph.D. degrees from the electrical and electronics engineering department of Bilkent University, Ankara, Turkey in 2010, 2012, and 2018, respectively. After studying the vector diffraction theory during his Ph.D. degree, he joined OzcanLab at UCLA in 2019. His work at UCLA focused on theoretical analysis on the generalization capacity and computational capabilities of diffractive optical networks. He is currently working as an optical design engineer at Aselsan.

**Aydogan Ozcan** is the Chancellor's Professor and the Volgenau Chair for Engineering Innovation at UCLA and an HHMI Professor with the Howard Hughes Medical Institute. He is also the Associate Director of the California NanoSystems Institute. He is elected Fellow of the National Academy of Inventors (NAI) and holds more than 50 issued/granted patents in microscopy, holography, computational imaging, sensing, mobile diagnostics, nonlinear optics, and fiber-optics and is also the author of one book and the co-author of more than 800 peer-reviewed publications in leading scientific journals/conferences. He received major awards, including the Presidential Early Career Award for Scientists and Engineers (PECASE), International Commission for Optics ICO Prize, Joseph Fraunhofer Award and Robert M. Burley Prize (Optica), SPIE Biophotonics Technology Innovator Award, Rahmi Koc Science Medal, SPIE Early Career Achievement Award, Army Young Investigator Award, NSF CAREER Award, NIH Director's New Innovator Award, Navy Young Investigator Award, IEEE Photonics Society Young Investigator Award and Distinguished Lecturer Award, National Geographic Emerging Explorer Award, National Academy of Engineering The Grainger

Foundation Frontiers of Engineering Award, and MIT's TR35 Award for his seminal contributions to computational imaging, sensing, and diagnostics. He is elected Fellow of Optica, AAAS, SPIE, IEEE, AIMBE, RSC, APS, and the Guggenheim Foundation, and is a Lifetime Fellow Member of Optica, NAI, AAAS, and SPIE. He is also listed as a Highly Cited Researcher by Web of Science, Clarivate.