Band-Structure-Engineered Electronic-Photonic Nonlinear Activation Functions

Zheheng Xu* and David Burghoff
Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA

(Received 20 June 2022; revised 5 October 2022; accepted 27 October 2022; published 14 December 2022)

Fast, sensitive, and compact devices that implement nonlinear activation functions are needed to form fully connected photonic neural networks (PNNs). However, even in highly nonlinear media, optical nonlinearities are relatively weak. We propose here a scheme for implementing nonlinear activation functions that relies on band-structure-engineered nanostructures. This scheme realizes the smallest possible hybrid optoelectronic approach, relying on fast electronic processes to implement nonlinearity instead of a true optical nonlinearity. Using well-established simplified density-matrix models, we demonstrate architectures that exhibit a low-intensity threshold of 3.5 μW along with a fast optical response of 10 ps in a relatively small linear footprint of 4 μm. We also show that PNN training performance is improved in handwritten pattern recognition when applying our simulated nonlinear activation function, indicating potential for creating deep fully connected PNNs.

DOI: 10.1103/PhysRevApplied.18.064038

I. INTRODUCTION

Deep learning algorithms have had a remarkable impact on many technologies [1]. One of the essential tools for deep learning, artificial neural networks, allow for any function to be learned given a sufficiently large network and training on a sufficiently large data set. Their power lies in parallel computations of two aspects: linear matrix-vector operations and nonlinear activation functions. Recent developments have greatly improved the performance of each, such as the use of graphical processing units for faster linear matrix multiplication [2] and the introduction of rectified linear units for nonlinear processing [3].

While traditional artificial neural networks use electronics as the information processing medium, photonic neural networks (PNNs) use photons to perform calculations [4]. PNNs could potentially be appealing for highly dense processing platforms due to their high power efficiency, low latency, and ultrawide bandwidths [5,6]. The linear computations of PNNs can be constructed with fairly simple optical elements, such as programmable Mach-Zehnder interferometers [7,8], wavelength division multiplexers [9], and diffractive optical elements [10,11]. Moreover, convolutional accelerators can, in principle, achieve speeds of 1 petaflops [12], and power efficiency of MZI-based passive nanophotonic circuits are reported at least 5 orders of magnitude better than conventional GPUs even for deep neural networks [7]. As more mature and systematic PNN models and fabrication processes may decrease the cost of large-scale photonic circuit integration [13–16], PNNs could become more comparable to electronic processors. However, PNNs still have several bottlenecks that have made them less competitive than electronic neural networks, particularly with respect to the nonlinear activation function. One of the main reasons for this is that optical nonlinearities [17,18] are intrinsically weak, making low-power operation challenging. Although it is possible to address this issue by enhancing nonlinearity in a resonant cavity or long waveguide, implementing high-intensity inputs, or other all-optical approaches, such as relying on free-carrier dispersion [19,20], phase change materials [21], or photonic lattices [22], the requirements on device footprint and energy consumption creates a severe speed-power-size trade-off. This limits the prospects for information processing and large-scale integration.

An alternative strategy relies on hybrid electronic-photonic approaches [23–27], which bypass this issue by relying on electronics to implement a nonlinearity. There are multiple ways of accomplishing this, but the most general implementation converts an optical signal into an electrical one, amplifies the electrical signal, and uses the amplified signal to drive a modulator of some kind. Provided the modulator response is nonlinear in the field, this can provoke a nonlinear response. One could also add additional nonlinearity through the use of conventional electronics. While these optical-to-electrical-to-optical approaches can, in principle, be compact and sensitive, they sacrifice most of the large bandwidth available to photons. In addition, they require sophisticated microwave engineering to properly extract and amplify the signal at high speeds. Meanwhile, their speeds are typically limited by the same mechanisms limiting conventional CMOS, which is already very developed.

*zxu7@nd.edu

From a fundamental standpoint, how small and fast can hybrid devices be made? Each of the core elements can be implemented using a two-level quantum system—detection by absorption, amplification by resonant tunneling, and modulation by the quantum-confined Stark effect [see Figs. 1(a) and 1(b)]—and so it stands to reason that only a few levels are needed to implement hybrid nonlinearity. In this work, we show that band-structure-engineered devices can act as integrated hybrid nonlinear activation functions, potentially acting as scalable drop-in elements for photonic neural networks.

The approach we consider here relies on intersubband (ISB) nanostructures and leverages the nanostructure's nonlinear electrical properties to modify its linear optical properties. As these devices do not require extraction of any electrical signal, they can, in principle, achieve operation speeds that are on par with the fastest electrical devices while achieving excellent sensitivity (μW thresholds), fast speeds (ps level), and in small footprints (μm^2 level). Based on numerical simulations of the ISB nanostructures using a periodic density matrix Schrödinger Poisson (DM SP) model, we demonstrate that such a nonlinear activation function has potential for creating deep fully connected PNNs. A large transmission contrast between light *on-off* states is found, while a low activation threshold and a relatively fast response time is achieved. Based on these results, a neural network is implemented that shows the capabilities of the ISB device as an elementwise nonlinear classifier. While the approach considered here relies on intersubband devices in established platforms and is therefore limited to operation in the midinfrared (analogous to quantum cascade lasers), similar principles can be used to make interband devices (analgous to interband cascade lasers), or intersubband devices based on high-barrier quantum wells.

II. BASIC PRINCIPLE

To illustrate the principle of how optoelectronic nonlinearity can arise in ISB nanostructures, we consider the simplest system that demonstrates this behavior: a three-level system. A schematic is shown in Fig. 1(b). The entire multiple quantum well system within one module is divided into a low-doped active region and a high-doped injector. The active region mainly involves intersubband transitions and consists of two aspects: an absorption transition where photons are absorbed and a resonant tunneling transition where electrons can tunnel. Because electronic transport is a function of photon absorption, this portion acts as a photodetector. In addition, if the main optical transition is diagonal, this portion of the structure will be sensitive to the field across it. In this sense, it acts as a modulator. Finally, the presence of a resonant tunneling stage acts to provide electrical gain. When it is properly biased, the current passing through the structure reaches a maximum; beyond this point it exhibits negative differential resistance (NDR). This structure can be repeated and inserted into a typical dielectric waveguide [see Fig. 1(c)].

To see how this structure gives rise to hybrid nonlinearity, assume that the injector region is heavily doped, so that its current-voltage relationship is approximately linear. The entire structure can then be treated as the active region (which exhibits NDR) and a load resistor connected in series [see Fig. 1(d)]. If the bias across the whole structure is chosen so that the load line just misses the peak of the off-state curve, the system will behave very differently depending on the incident intensity. At low illuminations, the active region is overbiased (point *B*), the optical transition is detuned, and the net absorption of the structure is low. When illumination exceeds the intensity threshold, the active region is properly biased (point *A*), electrons tunnel robustly and the absorption becomes high. Consequently, the transmission of the ISB nanostructure varies with the illumination's intensity, leading to a thresholding behavior and a nonlinear optical response of output intensity. Note that for different designs, the nanostructures could exhibit different activation behavior. For instance, the more detailed simulations in this work use a short injector with miniband resonant tunneling, which itself exhibits NDR. Thus, the absorption rate would reach a high level at low illumination while deceasing at high intensity. This hybrid approach achieves an effective nonlinear response without relying on nonlinear optics. Instead, it derives its nonlinear properties from an electrical nonlinearity present in an optical system.

The proposed ISB devices for PNN nonlinear activation have several unique features, chief among them being compactness. The detection comes from intersubband absorption, gain comes from resonant tunneling, and modulation comes from potential-tuned absorption. All of the elements in traditional hybrid approaches are still present, but within a few nanometers of each other. Another key feature of this approach is that its optical nonlinearity is induced by resonant tunneling, which is essentially the fastest known electrical process. The presence of NDR in resonant tunneling diodes can provide gain and nonlinearity up to terahertz frequencies [28], which is comparable with most PNNs using photodetection. Therefore, compared with optical-to-electrical-to-optical approaches [23,24,26,29], ISB nanostructures can work as highly integrated nonlinear activation circuits for ultrafast PNNs without separated photonics devices. This scheme provides more benefits over discrete optoelectronics since the signal is never extracted from the device, eliminating the sophisticated microwave engineering that would be required for high-speed operation. Our design also presents a convenient way for probing and *in situ* monitoring. Since this nanostructure is effectively a modified quantum well infrared photodetector, one could probe each neuron state in the system by simply

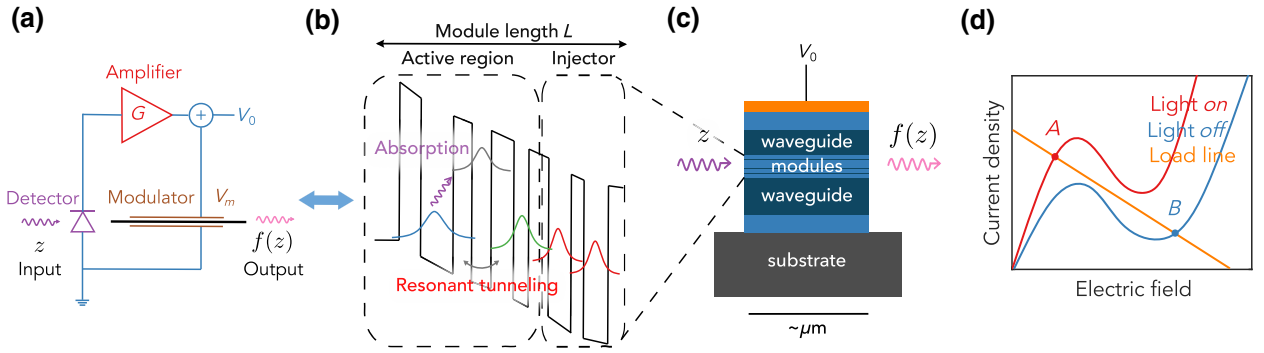

FIG. 1. (a) Simplest schematic representation of a hybrid electronic-photonic nonlinear activation function. (b) Three-level microscopic schematic for a nanostructure that implements the same functionality. (c) Vertical layout of an intersubband nanostructure implementing nonlinearity, showing the optoelectronic active region and waveguiding layers. The length of the device is on the order of micrometers. (d) Current-voltage relation of the active region under different illumination conditions. When biased, the current and voltage of active region are selected by the intersection with a finite impedance load line. At low illumination, the active region selects a high voltage. At high illumination, it selects a lower voltage.

measuring the current. One could similarly read the output of the system directly without any additional detectors, which is critical for efficiently training the network [30]. The nanostructures are reminiscent of quantum cascade laser (QCL) gain media and are highly compatible with QCL gain media. By growing gain media above or below the neuronal layers, regenerative gain could be added, which would be beneficial for deep PNNs [31].

These intersubband neuron devices have some design features in common with both midinfrared QCLs [32] and quantum well infrared photodetectors (QWIPs). Intersubband systems have very fast (approximately ps) scattering times, which is beneficial for making fast detectors [33] and frequency combs [34]. The waveguides and active region are on InP substrates and have periodic modules composed of ternary (In,Ga)As/(In,Al)As, one of the most well-established material systems available for intersubband photonics, with simulated waveguide losses of approximately 1 cm^{-1}. To implement a nonlinear optical transfer function, bias is added on the top of the device for threshold tuning and hysteresis reset (the highly doped substrate provides the ground). An incident photon generated by compact lasers such as QCLs is then shined into one side of the device, and the output is the light that has been modulated by the nonlinear absorption response. Of course, the major limitation of the structure considered here is that ternary (In,Ga)As/(In,Al)As can only be effectively designed in the midinfrared, as the barrier height is too low for the near-infrared telecommunication wavelength (1550 nm). We consider this system because it is the most well established and most well understood for band-structure engineering—to address this limitation, advances in material growth are needed (particular tall-barrier systems like the III nitrides).

III. THEORY

To demonstrate the efficacy of this approach, we use the well-established simplified density-matrix approach, which treats each subband as a single state and relies on effective scattering rates [35–38]. This allows us to calculate both the nonlinearity of our structure as well as its transient response. We use a nearest-neighbor (NN) tight-binding model, which allows states in adjacent modules to couple [39]. Each module contains N basis states that are calculated from the one-dimensional Schrödinger-Poisson equation:

$$-\frac{\hbar^2}{2}\frac{\partial}{\partial x}\frac{1}{m^*}\frac{\partial \psi}{\partial x} + V\psi = H\psi = E\psi, \quad x \in (0, L), \tag{1}$$

$$-\varepsilon_r \varepsilon_0 \frac{\partial^2 V}{\partial x^2} = e^2(n - n_D), \tag{2a}$$

$$V(0) = 0, \; V(L) = -eU, \tag{2b}$$

where ψ is the wave function, L indicates module length, $\hbar$ is the reduced Planck constant, m^* is the effective mass of an electron, U is the potential drop within one module, n_D is average doping density, ε_r is relative permittivity, and ε_0 is the permittivity of vacuum.

In order to statistically describe the interactions of these quantum states ψ, a simplified density-matrix model, which has been widely used in QCL simulations, is introduced. A general density matrix ρ is defined by $\rho = \sum_i \omega_i |\psi_i\rangle \langle\psi_i|$, where ω_i is the probability of the ith state. In a nearest-neighbor three-period system, the block matrix ρ and Hamiltonians can be expressed as

$$\rho = \begin{bmatrix} \rho_0 & \rho_{-1} & \rho_1 \\ \rho_1 & \rho_0 & \rho_{-1} \\ \rho_{-1} & \rho_1 & \rho_0 \end{bmatrix} \tag{3}$$

and

$$H = \begin{bmatrix} H_0 + eU & H_{-1} & H_1 \\ H_1 & H_0 & H_{-1} \\ H_{-1} & H_1 & H_0 - eU \end{bmatrix}, \tag{4}$$

where ρ_0 represents the density matrix in the center module and $\rho_1 = \rho_{-1}^{\dagger}$ represent the coherence of the center module with its neighbor. Each block contains N^2 matrix elements, as

$$\rho_0 = \begin{bmatrix} (\rho_0)_{11} & \cdots & (\rho_0)_{1N} \\ \vdots & \ddots & \vdots \\ (\rho_0)_{N1} & \cdots & (\rho_0)_{\mathrm{NN}} \end{bmatrix}. \tag{5}$$

The time evolution of the density matrix is given by the quantum Liouville equation:

$$\frac{\partial \rho}{\partial t} = \frac{1}{i\hbar}[H, \rho] + \frac{1}{i\hbar}[H', \rho] + \Gamma\rho, \tag{6}$$

where the first term describes the coherent transport of the system (corrected to account for the energy shift per module), the second term represents coherent interaction with the incident photon, and the third includes decoherence and scattering, obtained using the thermally averaged Fermi golden rule. It has a block matrix form of

$$\Gamma\rho = \begin{bmatrix} \Gamma\rho_0 & \Gamma_{||}\rho_{-1} & \Gamma_{||}\rho_1 \\ \Gamma_{||}\rho_1 & \Gamma\rho_0 & \Gamma_{||}\rho_{-1} \\ \Gamma_{||}\rho_{-1} & \Gamma_{||}\rho_1 & \Gamma\rho_0 \end{bmatrix}. \tag{7}$$

In the above expression, $\Gamma_{||}\rho$ contains only dephasing between the states of different periods, while $\Gamma\rho$ contains both intraperiod scattering and dephasing [40,41].

Each block can be described by

$$(\Gamma\rho)_{nm} = -\frac{1}{\hbar}\Gamma_{||nm}\rho_{nm}, \quad n \neq m, \tag{8a}$$

$$(\Gamma\rho)_{nn} = \frac{1}{\hbar}\left(\sum_{n\neq m}\Gamma_{mn}\rho_{mm} - \Gamma_n\rho_{nn}\right), \tag{8b}$$

where Γ_{mn} is the scattering between the mth state and nth state, Γ_n represents the total intersubband scattering rate, and $\Gamma_{||nm}$ represents the dephasing rate between the nth state and mth state, which has the following form:

$$\Gamma_n = \sum_{n\neq m}\Gamma_{nm}, \tag{9a}$$

$$\Gamma_{||nm} = \frac{1}{2}(\Gamma_{\mathrm{intra}} + \Gamma_m + \Gamma_n) + \frac{\hbar}{T_2^*}. \tag{9b}$$

Here, Γ_{intra} is the intrasubband scattering rate (which includes interface roughness and LO phonon scattering) and T_2^* is the pure dephasing that randomizes phase.

For simplification, the Liouville Eq. (6) can also be written in a linear system form with superoperators:

$$\frac{d\rho}{dt} = (L_C + L_S + L_D + L_{\mathrm{OD}})\,\rho. \tag{10}$$

This linearized equation (10) involves the coherent superoperator L_C, the scattering superoperator L_S, the dephasing superoperator L_D, and the optical drive superoperator L_{OD}.

To find steady-state solutions of the Liouville equation (10), the rotating-wave approximation is adopted. Diagonal elements of the density matrix indicate populations in different states, while off-diagonal elements denote coherence between states. After solving Eq. (1) numerically, a steady-state density matrix ρ is attained, and an alternative electron density can be found using

$$n(x) = n_D\rho(x, x), \tag{11}$$

where $\rho^{(0)}$ is the self-consistent steady-state density matrix under the initial environment. In this transient simulation algorithm illustrated as Fig. 2, the fast-changing density matrix would affect the slowly evolved electron density in Eq. (11). The potential distribution is then altered by Eq. (2a). Subsequently, a distorted band leads to varied wave functions by Eq. (1), generating an alternative Hamiltonian and coherent superoperator. This in turn changes the density matrix, and these effects continue until another balance is reached. Similarly, when the bias is varied in time the boundary condition of the Poisson equation can be allowed to be time-dependent; the absorption value is calculated at each time step and the entire response can be attained.

To implement the entire DM SP process, the potential is initialized to be linear under a constant bias. Iterations between Eqs. (1), (2a), (10), and (11) would yield a self-consistent steady-state DM SP solution. Although this calculation is fully quantum, to improve our understanding of device operation we also compute the semiclassical absorption rate using

$$\alpha = \bar{\Gamma}\frac{\Delta N e^2 f_{n\to m}\gamma(\nu)}{4m^* c n_r \varepsilon_0 L}, \tag{12a}$$

$$f_{n\to m} = \frac{2m^*(E_m - E_n)}{\hbar^2}\left|z_{n\to m}\right|^2, \tag{12b}$$

$$\Delta N = n_D(\rho_{nn} - \rho_{mm}), \tag{12c}$$

$$z_{n\to m} = \left\langle \psi_n \left| \hat{z} \right| \psi_m \right\rangle, \tag{12d}$$

where $\bar{\Gamma}$ is the mode confinement factor, $f_{n\to m}$ is the dimensionless oscillator strength, $z_{n\to m}$ is the dipole moment matrix element, E_m is the energy level of mth state, n_r is the refractive index, c is the speed of light, ν_0 is transition center frequency, and $\gamma(\nu)$ is the normalized lineshape, found from Eq. (9b).

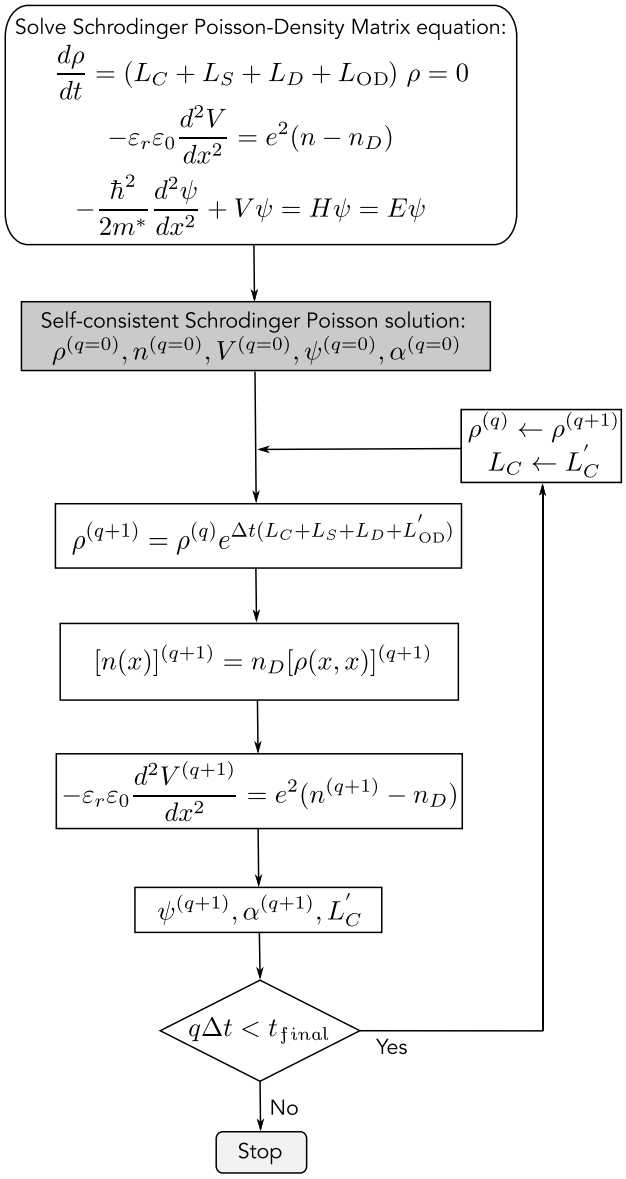

FIG. 2. Flow chart of the self-consistent transient DM SP algorithm, where q is the iteration number.

For transient simulations under time-varied illumination, we start by computing the self-consistent DM SP steady-state solution with an initial optical drive superoperator $L_{\rm OD}$. When the interaction superoperator changes from $L_{\rm OD}$ to $L'_{\rm OD}$ at $t_0 = 0$, all variables related to the density matrix reach an alternative equilibrium. The evolution of the density matrix can then be formally written in terms of matrix exponentiation as

$$\rho(t_0 + \Delta t) = \rho^{(0)} e^{(L_C+L_S+L_D+L'_{\rm OD})\Delta t}. \tag{13}$$

IV. DESIGN AND SIMULATION

To investigate the performance of hypothetical intersubband neurons, we design and simulate a device. We design a lattice-matched $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ ISB nanostructure with a device length $L_{\rm dev} = 1$ μm and a modal area of 2.3×5.2 μm^2. The ISB device with 57 modules is embedded in $In_{0.52}Al_{0.48}As$ waveguide grown on n-doped InP substrate for planar optical confinement. For such a waveguide, a simulation result shows a loss of 1.06 cm^{-1} and a fundamental mode confinement factor of 0.67 [see Fig. 3(a)]. The incident light has energy of 0.185 eV (wavelength of 6.7 μm) and interacts with the structure at optical powers ranging from 10^{-8} W to 10^{-3} W. The pure dephasing rate assumed for optical calculations and transport is 0.2 ps, while the device temperature is 300 K. In the device absorption simulation, we assume unity mode confinement factor for consistency. The position grid is 0.7 Å and the time step for transient simulations is 0.1 ps to ensure the convergence of the numerical results.

The periodic energy band profile under a constant bias of 2.3 V/μm is calculated in Fig. 3 from steady-state solution of the self-consistent DM SP model. The entire doping of the nanostructure is at a relatively high level (5×17 cm^{-3} for active region and 3.7×18 cm^{-3} for multiwell injector), which ensures that enough carriers are available to make absorption efficient, keeping the device footprint small and the impedance low. Due to miniband resonant tunneling (states in red shaded region), the structure possesses a NDR regime in the absence of light [see Fig. 4(a)]. In order to create appropriate NDR, barriers in the active region and multiwell injector should be carefully chosen. Too thick, and electrons do not efficiently tunnel into the next states and relax back down to the tunnel barriers, making the efficiency of the device low. Too thin, and the current-field relation will be less nonlinear, as for a two-level system current density is a Lorentzian function of the energy separation, with a linewidth proportional to the anticrossing strength [42]. In the absence of optical interaction, resonant tunneling within the active region is not sufficient. The absorption of the structure, mainly involving states $|1\rangle$ and $|2\rangle$, is tuned onto the laser frequency. As a result, the net absorption of the structure is high. When sufficient light is present (> 3.5 μW), a portion of electrons shared by main absorption states moves from left to right within the active region, causing population decrease and band distortion. The altered energy band thus aligns states that are responsible for resonant tunneling (especially states $|1\rangle$ and $|4\rangle$), creating a better channel for transport. In addition, the flow of electron density in the active region leads to a change of the wave-function shape of state $|2\rangle$ and $|3\rangle$, and the oscillator strength between the main absorption states decreases, making the ISB nanostructure more diagonal. Furthermore, the change of space charge would also detune the optical transition by varying the energy difference in the main absorption states. In summary, the combination of a diagonal structure, reduced population within the main absorption states, and frequency detuning can all suppress the absorption as

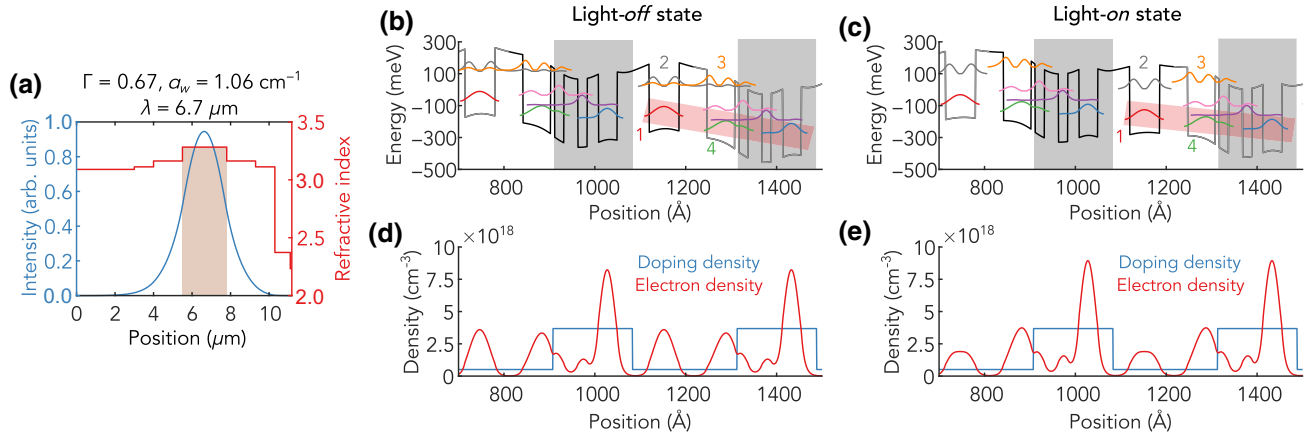

FIG. 3. (a) Fundamental mode profile of the design, with the shaded region indicating the active region of the device (57 modules). (b),(c) ISB nanostructure periodic energy band diagram and (d),(e) electron density distribution, with (5 μW) illumination (c),(e) and without illumination (b),(d). The gray shaded regions represent the injector, and the red shaded regions indicate resonant tunneling. The layer thicknesses, starting with the thick resonant tunneling barrier (where two modules connect), are **61.2**/60/**8.7**/20.3/**30.5**/20.3/**24.6**/40.4/**33**/**36.6**/65 Å. These thicknesses are chosen to be near an integer multiple of monolayers. The $In_{0.52}Al_{0.48}As$ barriers are shown in bold. Underlined layers are *n*-doped with a doping level of 3.7×10^{18} cm^{-3}, while other layers are *n*-doped with a doping level of 5×10^{17} cm^{-3}. Results are calculated self-consistently by the steady-state solution of the DM SP model with the boundary conditions that preserve charge neutrality.

light impinges on the structure [43], as shown in Fig. 4(b).

Figure 4(c) shows the potential distribution of three modules under different illumination conditions, showing that the potential distribution can change drastically between the light-*off* and light-*on* (5 μW) states. In the design, substantial difference of the doping level between the active region and the injector is adopted to enhance the effect. Even though the extra current is small, it causes the structure to lose electrical stability, abruptly changing the internal space charge. Within one module, this causes the bias to decrease in the active region while increasing the bias in the injector within one module. This potential drop redistribution results in insufficient bias on the active region, making the absorption decrease with higher optical power.

The absorption optical sensitivity of the ISB nanostructure under fixed bias is shown in Fig. 4(d). It exhibits behavior reminiscent of saturable absorption—transmitting high intensities while blocking low intensities, with a large transmission modulation of approximately 0.6 (corresponding to 4.8 dB/μm). Moreover, the absorption value remains nearly the same when the optical power is below or above the threshold; thus, it also acts as a step transmission function. In this ISB nanostructure design, a relatively small threshold at 3.5 μW is achieved due to the competing effects of the two different NDRs. In addition, the optical threshold can be tuned by selecting different constant biases, providing a degree of freedom for determining its nonlinear response and allowing for *in situ* optimization of the activation. As shown in the figure, the sensitivity at 310 K could be tuned to nearly the same one as 300 K by selecting different biases. Therefore, optimization of the bias would help the ISB device tolerate temperature changes, or other variations like doping and well width fluctuations.

The transient evolution of the transmission is shown in Fig. 5. When the system is in a high-transmission state and the optical intensity of 0.1 mW turns off, the system responds by transiting to a low-transmission state. This occurs in just 10 ps, which would be difficult to accomplish with pure electronics. However, we find that this design also exhibits a hysteresis, caused by the space charge becoming trapped on one side of the injection barrier. Conceptually, this is similar to what occurs in high-sensitivity photodetectors, which usually need to be quenched once they have fired. While such hysteresis effects could be used to store information [44] and to make recurrent neural networks, it is not ideal for the implementation of straightforward networks. To bypass this issue, we reset the device from light-*off* state to light-*on* state by applying a 40-ps ramp to the applied bias to rearm it. (Note that this scheme does not require information processing: rearming could be performed at very high speeds and at a constant rate, such as by a single global clock. In addition, the entire neuron network could be rearmed at the same time by only a single voltage modulator.) Once the device has been reset, it is free to fire once again. Note also that fast transmission change also occurs when the light turns on (without hysteresis), although the contrast is reduced.

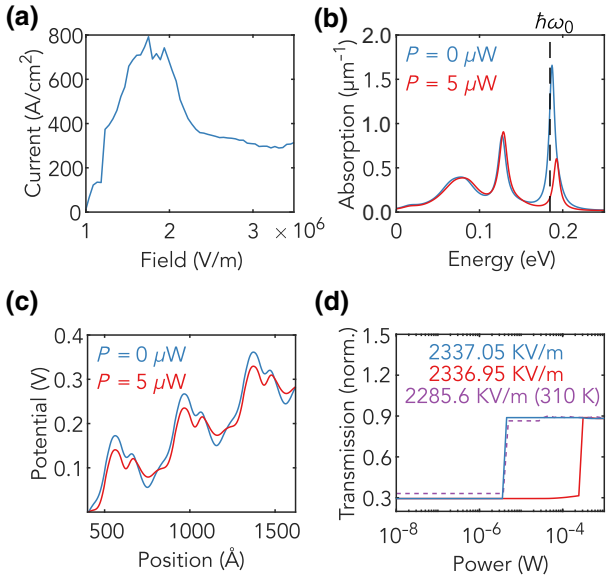

FIG. 4. (a) Simulated current-field relation without illumination. (b) Absorption (in the modules) and (c) potential distribution (of three modules) comparison between ISB light-*on* state and light-*off* state under constant applied field. (d) Transmission as a function of input optical power and structure bias. The transmission is effectively a step function, whose threshold can be tuned by choice of structure bias. The device length is $L_{\text{dev}} = 1$ μm, the modal area is 2.3×5.2 μm^2, and the temperature is 300 K except when stated otherwise. All data is taken from steady-state simulations.

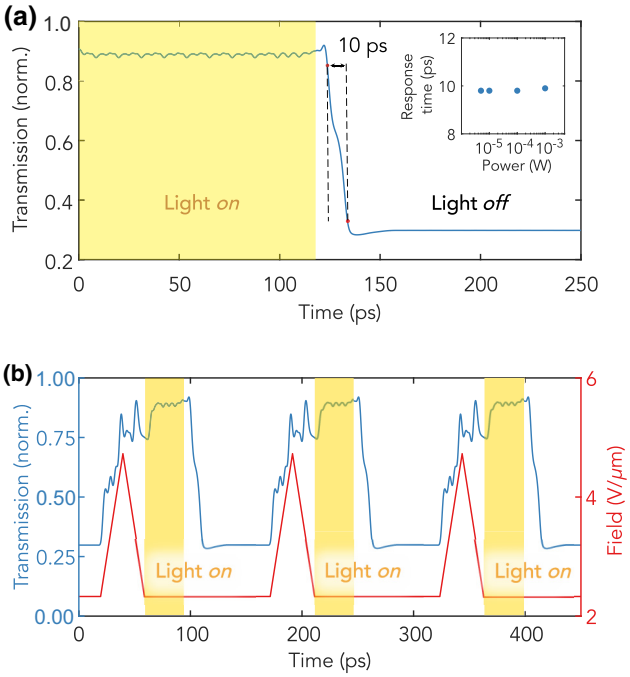

FIG. 5. Transmission versus time during a *on-off* switching event. (a) Response time of transmission with a change of lighting condition. The light intensity during the transient evolution is 0.1 mW, and response time is defined as the time duration between 10% and 90% of the stable levels after switching intensity. Inset shows response time as a function of input optical power under light-*on* condition. (b) Periodic transmission-time characteristics (using a 40-ps ramp to reset hysteresis from light-*off* to light-*on* states).

V. HANDWRITTEN DIGIT RECOGNITION

Next, we demonstrate the efficacy of our devices as a nonlinear classifier in a simulated network. In the ISB nanostructure, transmission as a function of input power resembles a step function, creating the nonlinearity needed for PNNs. Our simulation results demonstrate that these devices can have small footprints, low optical thresholds, and relatively short latencies. Therefore, ISB devices represent a promising direction for nonlinear activation function in PNNs operated at high speeds and low power. The activation function according to the simulated transmission optical sensitivity, which describes the normalized output signal intensity as a function of normalized input signal intensity, is shown in Fig. 6(c). The nonlinear activation behaves like a modified parametric rectified linear unit, with a suppressed transmission for inputs with low intensity and a large transmission for inputs with intensity above threshold [23]. It is also an odd function since the difference between signals with positive and negative intensity is only reflected by their phases [45]. In addition, the ISB nanostructure is basically a square-law device and does not need to be phase coherent.

For the demonstration of the PNN-based handwritten-digit recognition task, a feed-forward two-layer shallow network is configured using handwritten digits obtained from the Modified National Institute of Standards and Technology (MNIST) database, which is one of the most commonly used datasets in machine learning [46]. It contains 60 000 training images and 10 000 testing images with 28×28 grayscale pixels, labeled by numbers 0–9. The grayscale value of each pixel is normalized into the range of $[0, 1]$ for fitting the normalized input intensity.

The structure of a fully connected feed-forward optical neuron network in this application is depicted in Fig. 6(a). The entire PNN architecture consists of 784 inputs, corresponding to 28^2 real pixel value, and ten final outputs, corresponding to ten digits. There are $N_n = 40$ neurons in the hidden layer and ten neurons in the output layer. In each layer, the information vector is multiplied by a weight matrix and then processed by an elementwise activation function to generate outputs. In the demonstration, a 4-μm-long device is chosen since it has the largest amplitude modulation depth, which could offer the best activation nonlinearity [see Fig. 6(b)]. We also assume that the large number of optical inputs is achieved by high-power laser with power splitters [47] or an array of QCLs [48], and the PNN operates under optical pulse width larger than

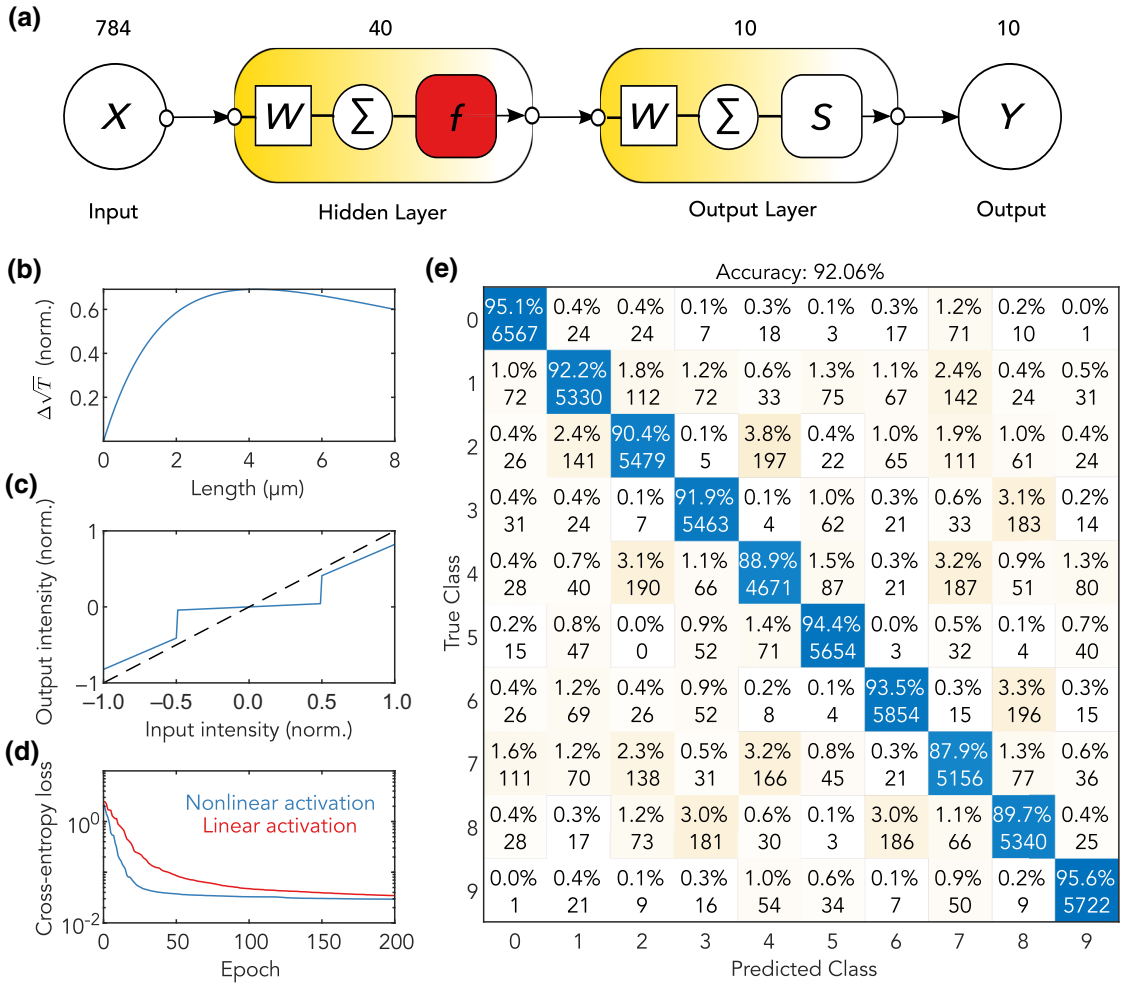

FIG. 6. Overview of the ISB nanostructure applied in a machine learning task. (a) General two-layer PNN architecture for image recognition. (b) Modulation depth as a function of device length. (c) Nonlinear activation function based on the ISB nanostructure (blue lines) and linear activation function (dashed lines). (d) Cross-entropy loss performance with and without activation function. (e) Confusion matrix for trained PNN with activation functions. The number of observations and the corresponding percentages are shown in each cell.

optical response time of the device to reach the light-*on* steady state. A comparison between a linear classifier and a nonlinear classifier based on ISB nanostructures (ISB nonlinear activation function) is carried out in a hidden layer. The loss function is computed by *softmax*, which is commonly used for multiclass image classification [49]. It normalizes the intensity outputs of the PNN into a probability distribution over predicted classes, and the corresponding performance function is cross-entropy loss. After adequate data feeding and prediction error optimization by backpropagation, trained PNN could perform image recognition tasks.

During each training epoch, training data is divided randomly into training and validation subsets with a ratio of 4 : 1. After feeding the network with the training set, the remaining testing images are used to compute the accuracy and confusion matrices. The training performance comparison is shown in Fig. 6(d). It can be observed that compared with a linear classifier, the nonlinear activation function improves PNN performance by increasing training speed and decreasing errors. The confusion matrix is also shown in Fig. 6(e). The final accuracy of PNN with the ISB nonlinear activation function is 92% for 200 epochs of training (the linear one is 90.8%). Moreover, other evaluations of the nanostructure's performance are also carried out to show the practicality of ISB activation in PNNs.

To evaluate the performance of this network, we follow the convention in Ref. [23] and ignore electrical control lines and coupling waveguides. The estimated footprint of the nanostructures in the PNN is $A = LN_nL_{\mathrm{dev}}W_{\mathrm{dev}}$, where $L = 1$ is the number of layers and W_{dev} is the device width. The estimated energy consumption for the nonlinear activation can be expressed as $P_n = LN_nJVL_{\mathrm{dev}}W_{\mathrm{dev}}H_{\mathrm{dev}}$, where J is the current density, V is the bias applied on

TABLE I. Comparison between the ISB device approach and optoelectronic approaches in PNN nonlinearity. The activation functions are compared in terms of main figures of merit: transmission modulation (dB/μm), activation threshold (μW), latency, linear footprint (μm), working wavelength (μm), and energy consumption (μW).

Approach	Transmission modulation	Activation threshold	Latency	Linear footprint	Wavelength	Energy consumption
This work	4.8 dB/μm	3.5 μW	10 ps	4 μm	6.7 μm	562 μW
Ref. [23]	$>$ 0.2 dB/μm	100 μW	120 ps	$<$ 60 μm	...	$>$ 10 000 μW
Ref. [24]	$>$ 1 dB/μm	...	160 ps	11.5 μm	1.545 μm	$>$ 1000 μW
Ref. [26]	0.1 dB/μm	$>$ 30 μW	50 ps	$>$ 25 μm	...	...
Ref. [50]	0.132 dB/μm	$>$ 40 μW	1000 ps	15 μm	1.55 μm	$>$ 1000 μW
Ref. [51]	1 dB/μm	$>$ 10 μW	40 ps	5 μm	1.55 μm	1500 μW
Ref. [52]	1.2 dB/μm	...	300 s	5 μm	1.55 μm	$>$ 0.2 μW
Ref. [53]	...	...	10 ms	$>$ 2 μm	1.55 μm	$>$ 0.15 μW

the device, and H_{dev} is the height of the device. The energy consumption during the rearming process could be express as $P_{\mathrm{re}} = (\int_{t_1}^{t_2} LN_n J(t)V(t)L_{\mathrm{dev}}W_{\mathrm{dev}}H_{\mathrm{dev}}dt)/(t_2 - t_1)$, where $t_2 - t_1$ is the time for rearming process. By taking the same duration for both nonlinear activation and rearming process, the average overall power consumption is $(P_n + P_{\mathrm{re}})/2 + P_{\mathrm{ac}}$, where P_{ac} is the optical threshold power needed for the PNN. For a device dimension of $4 \times 5.2 \times 2.3$ μm, the single activation performance with comparisons of other optoelectronic methods is also shown in Table I. It is noted that the physical footprint of the activation function with 40 neurons in the demonstration is only 8.3×10^{-4} mm^2 per layer with an average total power consumption of approximately 22.48 mW under operation, which is superior to other optoelectronic methods. Moreover, the efficiency of the ISB nanostructures could be greatly improved. For instance, one could add wells that suppress leakage using shorter modules or rely on higher-barrier material systems. Similar nanostructures could be the scalable elements that allow for deep photonic neural networks with even millions of neurons.

VI. DISCUSSION

It is worthwhile to emphasize that the origin of the apparent nonlinearity in these devices is not due to optical nonlinearity, it is due to band bending arising from coherent population transfer. Because this effect is ultimately electronic, this allows a substantial transmission change at micron-scale lengths, much higher than similar-sized devices relying on giant intersubband $\chi^{(3)}$ nonlinearities at the same power. For example, though it is possible to achieve Kerr nonlinearities of $n_2 \sim 10^5 - 10^6$ nm^2/W in intersubband structures [54,55], at powers of 5 μW, if the Kerr effect is used in conjunction with an interferometer to create a modulator, such a device would require that the path be $L_{\mathrm{ISB}} = 4.59$ m long (assuming a Kerr nonlinearity of 1×10^6 nm^2/W). In this regard, the nonlinearity can be considered an ultrafast electronic nonlinearity like that present in resonant tunneling diodes [56]. In these devices, the ultimate speed is limited by the subpicosecond phonon-scattering relaxation times; similarly, the ISB devices could also respond to changes in incident light on picosecond time scales. Moreover, the nonlinear activation has a hysteresis, which means the device could be exploited to behave similar to optoelectronic memristors [57–59]. In principle, the same band structure could be used to act as nonlinear activation functions, as optical detectors or amplifiers, and as optical memory storage units, which would be beneficial for large-scale integration.

As for the activation threshold, its origin is more complicated. As this device combines features of the three most-common intersubband devices, the design space is complex and has significant room for improvement. For example, realistic designs typically have NDR in both the active region and the injector, which must be accounted for. Moreover, by shifting the absorption frequency, the structure can be made to act in the reverse mode of nonlinear operation, transmitting low intensities while blocking high intensities. Therefore, most figures of merit of ISB devices could be further improved or modified by ISB structure parameter optimization with different injection schemes. In addition, compared with electro-optic nonlinear activation functions where light is tapped, detected, and used to drive an intensity modulator, all these elements of our ISB device—photon detection, high-speed gain, and electroabsorption—are effectively contained within a single nanostructure. Thus the main challenge for ISB activation in deep PNNs would not be the device themselves, but rather the large diameters of the waveguides and the larger optical losses. In our structure, due to the high leakage currents there is an estimated dc power consumption of 384 μW for a single activation function in PNN, which could be further reduced by managing the leakage mechanism of the system [60].

Although we consider only ISB nonlinear activation functions operating in the midinfrared using (In,Ga)As/(In,Al)As system, one could extend the same principles to the near-infrared by using GaN/AlN system, where essentially every metric—device size,

power consumption, nonlinear threshold, response time, etc.—improve drastically, due to the increased photon energy. For example, the QWIP-like transport mechanism experiences electron leakage similar to thermionic emission [61]. At room temperature and a wavelength of 6.7 μm, this leakage mechanism dominates, resulting in an unavoidable dark current that scales with $\exp(-\hbar\omega_0/kT)$. Therefore, higher optical frequency in ISB activation devices would not only reduce the power dissipation exponentially, but it would also decrease the dark current noise (related to the square root of the leakage current [62]). The noise improvement could further lead to a smaller activation threshold in practical PNN applications of ISB devices. In addition, GaN/AlN-based devices are attractive for terahertz modulation frequencies due to their extremely short absorption recovery times [63]. Other advantages of the GaN/AlN system for ISB nonlinear activation functions include the feasibility of low-loss integrated photonic circuits at near-infrared wavelengths and improved thermal robustness of devices. However, the GaN/AlN system is not as well understood due to its less-mature growth technology (for example, the effect of built-in fields and the ultimate interface roughness that can be achieved). Given the recent development of room-temperature high-frequency GaN/AlN resonant tunneling diodes [64,65], quantum cascade detectors [66], and QWIPs [67], the realization of near-infrared intersubband nonlinear activation devices in the future could potentially revolutionize deep photonic neural networks.

VII. CONCLUSION

In conclusion, we introduce a strategy for achieving nonlinear optical activation functions based on band-structure-engineered nanostructures. Our simulations revealed that the designed ISB nanostructures are capable of high-speed nonlinear processing in deep PNNs. In contrast to standard optoelectronic approaches, this ISB architecture leverages the nanostructure's nonlinear electrical properties to modify its linear optical properties. Therefore, this approach could achieve a low activation threshold around 3.5 μW with fast response of 10 ps while maintaining a single linear footprint of 4 μm, which is much smaller than previously proposed schemes. Finally, based on numerical simulations of the ISB nanostructures, we demonstrate such a nonlinear activation function enhances PNNs' performance on the benchmark task of handwritten-number recognition from the MNIST dataset. This approach has significant potential for the creation of deep fully connected PNNs.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2046772. In addition, D.B. acknowledges support from AFOSR Grant No. FA9550-20-1-0192 and ONR Grant No. N00014-21-1-2735; this research is funded in part by the Gordon and Betty Moore Foundation through Grant No. GBMF11446 to the University of Notre Dame to support the work of D.B.

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature **521**, 436 (2015).
[2] V. K. Pallipuram, M. Bhuiyan, and M. C. Smith, A comparative study of GPU programming models and architectures using neural networks, J. Supercomput. **61**, 673 (2012).
[3] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, in *2009 IEEE 12th International Conference on Computer Vision* (IEEE, Kyoto, Japan, 2009), p. 2146.
[4] F.-C. F. Tsai, C. J. O'Brien, N. S. Petrović, and A. D. Rakić, Analysis of optical channel cross talk for free-space optical interconnects in the presence of higher-order transverse modes, Appl. Opt. **44**, 6380 (2005).
[5] H. J. Caulfield, J. Kinser, and S. K. Rogers, Optical neural networks, Proc. IEEE **77**, 1573 (1989).
[6] P. Ghelfi, F. Laghezza, F. Scotti, G. Serafino, A. Capria, S. Pinna, D. Onori, C. Porzi, M. Scaffardi, A. Malacarne, Valeria Vercesi, Emma Lazzeri, Fabrizio Berizzi, and Antonella Bogoni, A fully photonics-based coherent radar system, Nature **507**, 341 (2014).
[7] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and Marin Soljačić, Deep learning with coherent nanophotonic circuits, Nat. Photonics **11**, 441 (2017).
[8] H. Zhang, M. Gu, X. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. 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**, 1 (2021).
[9] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, Parallel convolutional processing using an integrated photonic tensor core, Nature **589**, 52 (2021).
[10] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, All-optical machine learning using diffractive deep neural networks, Science **361**, 1004 (2018).
[11] 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).
[12] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, Arnan Mitchell, and David J. Moss, 11 tops photonic convolutional accelerator for optical neural networks, Nature **589**, 44 (2021).
[13] Q. Zhang, Z. Xing, and D. Huang, Implementation of pruned backpropagation neural network based on photonic integrated circuits, Photonics **8**, 363 (2021).
[14] D. J. Moss, R. Morandotti, A. L. Gaeta, and M. Lipson, New CMOS-compatible platforms based on silicon nitride

and Hydex for nonlinear optics, Nat. Photonics **7**, 597 (2013).

[15] N. M. Fahrenkopf, C. McDonough, G. L. Leake, Z. Su, E. Timurdogan, and D. D. Coolbaugh, The AIM photonics MPW: A highly accessible cutting edge technology for rapid prototyping of photonic integrated circuits, IEEE. J. Sel. Top. Quantum Electron. **25**, 1 (2019).

[16] T. Heuser, J. Große, A. Kaganskiy, D. Brunner, and S. Reitzenstein, Fabrication of dense diameter-tuned quantum dot micropillar arrays for applications in photonic information processing, APL Photonics **3**, 116103 (2018).

[17] B. Wu, H. Li, W. Tong, J. Dong, and X. Zhang, Low-threshold all-optical nonlinear activation function based on a Ge/Si hybrid structure in a microring resonator, Opt. Mater. Express **12**, 970 (2022).

[18] C. Huang, S. Fujisawa, T. F. de Lima, A. N. Tait, E. C. Blow, Y. Tian, S. Bilodeau, A. Jha, F. Yaman, H.-T. Peng, Hussam G. Batshon, Bhavin J. Shastri, Yoshihisa Inada, Ting Wang, and Paul R. Prucnal, A silicon photonic–electronic neural network for fibre nonlinearity compensation, Nat. Electron. **4**, 837 (2021).

[19] A. Jha, C. Huang, and P. R. Prucnal, Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics, Opt. Lett. **45**, 4819 (2020).

[20] H. Li, B. Wu, W. Tong, J. Dong, and X. Zhang, All-optical nonlinear activation function based on germanium silicon hybrid asymmetric coupler, IEEE Journal of Selected Topics in Quantum Electronics (early access) (2022).

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

[22] A. V. Pankov, I. D. Vatnik, and A. A. Sukhorukov, Optical Neural Network Based on Synthetic Nonlinear Photonic Lattices, Phys. Rev. Appl. **17**, 024011 (2022).

[23] I. A. 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 (2019).

[24] 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).

[25] A. Totovic, C. Pappas, M. Kirtas, A. Tsakyridis, G. Giamougiannis, N. Passalis, M. Moralis-Pegios, A. Tefas, and N. Pleros, WDM equipped universal linear optics for programmable neuromorphic photonic processors, Neuromorphic Comput. Eng. **2**, 024010 (2022).

[26] 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 (2019).

[27] H.-T. Peng, G. Angelatos, T. F. de Lima, M. A. Nahmias, A. N. Tait, S. Abbaslou, B. J. Shastri, and P. R. Prucnal, Temporal information processing with an integrated laser neuron, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2019).

[28] R. Izumi, S. Suzuki, and M. Asada, in *2017 42nd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)* (IEEE, Cancun, Mexico, 2017), p. 1.

[29] M. M. P. Fard, I. A. Williamson, M. Edwards, K. Liu, S. Pai, B. Bartlett, M. Minkov, T. W. Hughes, S. Fan, and T.-A. Nguyen, Experimental realization of arbitrary activation functions for optical neural networks, Opt. Express **28**, 12138 (2020).

[30] 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 (2018).

[31] F. Ashtiani, A. J. Geers, and F. Aflatouni, An on-chip photonic deep neural network for image classification, Nature **606**, 501 (2022).

[32] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, Quantum cascade laser, Science **264**, 553 (1994).

[33] D. Palaferri, Y. Todorov, A. Bigioli, A. Mottaghizadeh, D. Gacemi, A. Calabrese, A. Vasanelli, L. Li, A. G. Davies, E. H. Linfield, Filippos Kapsalidis, Mattias Beck, Jérôme Faist, and Carlo Sirtori, Room-temperature nine-μm-wavelength photodetectors and GHz-frequency heterodyne receivers, Nature **556**, 85 (2018).

[34] A. Hugi, G. Villares, S. Blaser, H. C. Liu, and J. Faist, Mid-infrared frequency comb based on a quantum cascade laser, Nature **492**, 229 (2012).

[35] B. A. Burnett and B. S. Williams, Density matrix model for polarons in a terahertz quantum dot cascade laser, Phys. Rev. B **90**, 155309 (2014).

[36] A. Pan, B. A. Burnett, C. O. Chui, and B. S. Williams, Density matrix modeling of quantum cascade lasers without an artificially localized basis: A generalized scattering approach, Phys. Rev. B **96**, 085308 (2017).

[37] R. Terazzi and J. Faist, A density matrix model of transport and radiation in quantum cascade lasers, New J. Phys. **12**, 033045 (2010).

[38] S. Kumar and Q. Hu, Coherence of resonant-tunneling transport in terahertz quantum-cascade lasers, Phys. Rev. B **80**, 245316 (2009).

[39] A. Demić, Z. Ikonić, R. W. Kelsall, and D. Indjin, Density matrix superoperator for periodic quantum systems and its application to quantum cascade laser structures, AIP. Adv. **9**, 095019 (2019).

[40] T. V. Dinh, A. Valavanis, L. J. M. Lever, Z. Ikonić, and R. W. Kelsall, Extended density-matrix model applied to silicon-based terahertz quantum cascade lasers, Phys. Rev. B **85**, 235427 (2012).

[41] A. Demić, A. Grier, Z. Ikonić, A. Valavanis, C. A. Evans, R. Mohandas, L. Li, E. H. Linfield, A. G. Davies, and D. Indjin, Infinite-period density-matrix model for terahertz-frequency quantum cascade lasers, IEEE Trans. Terahertz Sci. Technol. **7**, 368 (2017).

[42] B. S. Williams, Terahertz quantum-cascade lasers, Nat. Photonics **1**, 517 (2007).

[43] S. Kumar, Q. Hu, and J. L. Reno, 186 K operation of terahertz quantum-cascade lasers based on a diagonal design, Appl. Phys. Lett. **94**, 131105 (2009).

[44] M. I. Stockman, L. N. Pandey, L. S. Muratov, and T. F. George, Intersubband optical bistability induced by resonant tunneling in an asymmetric double quantum well, Phys. Rev. B **48**, 10966 (1993).

[45] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and Marin Soljacic, Supplementary information:

Deep learning with coherent nanophotonic circuits, Nat. Photonics **11**, 441 (2017).

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).

[47] X. Xiao, M. B. On, T. Van Vaerenbergh, D. Liang, R. G. Beausoleil, and S. B. Yoo, Large-scale and energy-efficient tensorized optical neural networks on III–V-on-silicon MOSCAP platform, APL Photonics **6**, 126107 (2021).

[48] W. Zhou, S. Slivken, and M. Razeghi, Phase-locked, high power, mid-infrared quantum cascade laser arrays, Appl. Phys. Lett. **112**, 181106 (2018).

[49] W. Liu, Y. Wen, Z. Yu, and M. Yang, in *International Conference on Machine Learning* (PMLR, New York, USA, 2016), p. 507.

[50] R. Amin, J. K. George, H. Wang, R. Maiti, Z. Ma, H. Dalir, J. B. Khurgin, and V. J. Sorger, An ITO–graphene heterojunction integrated absorption modulator on Si-photonics for neuromorphic nonlinear activation, APL Photonics **6**, 120801 (2021).

[51] R. Amin, J. George, S. Sun, T. Ferreira de Lima, A. N. Tait, J. 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).

[52] C. Hoessbacher, Y. Fedoryshyn, A. Emboras, A. Melikyan, M. Kohl, D. Hillerkuss, C. Hafner, and J. Leuthold, The plasmonic memristor: A latching optical switch, Optica **1**, 198 (2014).

[53] A. Emboras, I. Goykhman, B. Desiatov, N. Mazurski, L. Stern, J. Shappir, and U. Levy, Nanoscale plasmonic memristor with optical readout functionality, Nano Lett. **13**, 6151 (2013).

[54] J. Bai and D. Citrin, Enhancement of optical Kerr effect in quantum-cascade lasers with multiple resonance levels, Opt. Express **16**, 12599 (2008).

[55] P. Friedli, H. Sigg, B. Hinkov, A. Hugi, S. Riedi, M. Beck, and J. Faist, Four-wave mixing in a quantum cascade laser amplifier, Appl. Phys. Lett. **102**, 222104 (2013).

[56] T. Maekawa, H. Kanaya, S. Suzuki, and M. Asada, Oscillation up to 1.92 THz in resonant tunneling diode by reduced conduction loss, Appl. Phys. Express **9**, 024101 (2016).

[57] T.-Y. Wang, J.-L. Meng, Q.-X. Li, Z.-Y. He, H. Zhu, L. Ji, Q.-Q. Sun, L. Chen, and D. W. Zhang, Reconfigurable optoelectronic memristor for in-sensor computing applications, Nano Energy **89**, 106291 (2021).

[58] L. Hu, J. Yang, J. Wang, P. Cheng, L. O. Chua, and F. Zhuge, All-optically controlled memristor for optoelectronic neuromorphic computing, Adv. Funct. Mater. **31**, 2005582 (2021).

[59] Z.-D. Luo, X. Xia, M.-M. Yang, N. R. Wilson, A. Gruverman, and M. Alexe, Artificial optoelectronic synapses based on ferroelectric field-effect enabled 2D transition metal dichalcogenide memristive transistors, ACS Nano **14**, 746 (2019).

[60] H. T. Miyazaki, T. Mano, T. Kasaya, H. Osato, K. Watanabe, Y. Sugimoto, T. Kawazu, Y. Arai, A. Shigetou, T. Ochiai, Yoji Jimba, and Hiroshi Miyazaki, Synchronously wired infrared antennas for resonant single-quantum-well photodetection up to room temperature, Nat. Commun. **11**, 1 (2020).

[61] H. Liu, A. Steele, M. Buchanan, and Z. Wasilewski, Dark current in quantum well infrared photodetectors, J. Appl. Phys. **73**, 2029 (1993).

[62] B. Levine, A. Zussman, J. Kuo, and J. De Jong, 19μm cutoff long-wavelength GaAs/Al_xGa_{1-x} as quantum-well infrared photodetectors, J. Appl. Phys. **71**, 5130 (1992).

[63] N. Iizuka, K. Kaneko, and N. Suzuki, Near-infrared intersubband absorption in GaN/AlN quantum wells grown by molecular beam epitaxy, Appl. Phys. Lett. **81**, 1803 (2002).

[64] T. A. Growden, D. F. Storm, E. M. Cornuelle, E. R. Brown, W. Zhang, B. P. Downey, J. A. Roussos, N. Cronk, L. B. Ruppalt, J. G. Champlain, Paul R. Berger, and David J. Meyer, Superior growth, yield, repeatability, and switching performance in GaN-based resonant tunneling diodes, Appl. Phys. Lett. **116**, 113501 (2020).

[65] W.-D. Zhang, T. Growden, D. Storm, D. Meyer, P. Berger, and E. Brown, Investigation of switching time in GaN/AlN resonant tunneling diodes by experiments and p-spice models, IEEE Trans. Electron Devices **67**, 75 (2019).

[66] P. Quach, S. Liu, A. Jollivet, D. Wang, J. Cheng, N. Isac, S. Pirotta, D. Bouville, S. Sheng, A. Imran, L. Chen, D. Li, X. T. Zheng, Y. X. Wang, Z. X. Qin, M. Tchernycheva, F. H. Julien, B. Shen, and X. Q. Wang, A GaN/AlN quantum cascade detector with a broad response from the mid-infrared (4.1 μm) to the visible (550 nm) spectral range, Appl. Phys. Lett. **116**, 171102 (2020).

[67] P. M. Mensz, B. Dror, A. Ajay, C. Bougerol, E. Monroy, M. Orenstein, and G. Bahir, Design and implementation of bound-to-quasibound GaN/AlGaN photovoltaic quantum well infrared photodetectors operating in the short wavelength infrared range at room temperature, J. Appl. Phys. **125**, 174505 (2019).