Machine Vision with InP based Floating-gate Photo-field-effective Transistors for Color-mixed Image Recognition

Jun Tao, Juan Sanchez Vazquez, Hyun Uk Chae, Ragib Ahsan, Rehan Kapadia

***Abstract*—The success of artificial neural networks (ANNs) in machine vision techniques has driven hardware researchers to explore more efficient computing elements for energy-expensive operations such as vector-matrix multiplication (VMM). In this work, InP-based floating-gate photo-field-effective transistors (FG-PFETs) are demonstrated as computing elements that integrate both photodetection and initial signal processing at the sensor level. These devices are fabricated from semiconductor channels grown via a back-end CMOS compatible templated-liquid phase (TLP) approach. Individual devices are shown to exhibit programmable responsivity, mimicking the effect of a synapse connecting the photodetector to a neuron. Using these devices, a simulated optical neural network (ONN) where the experimentally measured performance of FG-PFETs is used as an input shows excellent image recognition accuracy for color-mixed handwritten digits.**

***Index Terms*— Machine vision, optical neural network, phototransistors, indium phosphide.**

I. Introduction

In the past few decades, machine vision has become a key component of various intelligent systems, including autonomous vehicles, automated fabrication systems, and robotics [1-4]. Due to the variety and sophistication of algorithms for machine vision, and in particular artificial neural networks (ANN) and convolutional neural networks (CNN), image recognition accuracy for well-trained machines can surpass human beings in specific tasks [5, 6]. However, this is typically achieved through large sets of training data, long training times, considerable hardware, and a optimization of parameters for the task at hand. To continue progress, development of new device structures and circuit architectures are heavily explored with the aim of reducing energy consumption, improving speed, and maintaining overall accuracy.

Many hardware level solutions based on CMOS technology and emerging devices built with advanced materials have been proposed to emulate biological neural or synaptic behaviors and boost the computation speed of ANNs [7, 8]. At the device level, resistive random access memory (RRAM) [9, 10], phase change memory (PCM) [8, 11], ferroelectric field effective transistor (FeFET) [12], ion-based electrolyte-gated transistor[13-15], and spin transfer torque magnetoresistive random access memory (STT-MRAM) [16] have been heavily studied. Most of the key computationally required synaptic behaviors like short- and long-term plasticity, spike-number-dependent plasticity, and spike-timing-dependent plasticity are successfully emulated [17-19]. Importantly, chip level acceleration of the vector matrix multiplication (VMM) operation has also been demonstrated by several research groups [20-23]. In addition, many novel computing elements coupled with biometric sensing functions were reported[14, 24, 25]. Among these, photonic synapses combine sensing and processing in a single device, which eliminates the additional sensing element, possessing the advantages of low energy consumption and large bandwidth[26, 27]. Conventionally, the image recognition algorithms with ANNs require digitized image inputs, a conversion that consumes energy and process time. But the characteristics of photonic synapses potentially can solve the bottleneck here. One such approach was a WSe_2 based two-dimensional (2D) semiconductor photodiode array reported to be able to sense and classify the image projected to the chip and demonstrated ultrafast processing time [1]. One challenge faced by such approaches however are the manufacturability and scalability of such processes, which require growth and transfer, as well as the relatively weak light absorption in thin 2-D materials. Additionally, the synaptic weight was stored off chip and used to modify the voltage applied to the gate of the device, an approach which is not feasible on a chip as it would require a tremendous number of voltages.

To approach this challenge, we use a scalable III-V growth approach that potentially enables monolithic integration with Si CMOS circuits and a device architecture that contains memory to store the responsivity values required. Specifically, we report an indium phosphide (InP) based floating-gate photo-field-

Manuscript received Dec 1, 2021. This work was supported in part by the U.S. National Science Foundation with award no. 1610604 and 2004791. (*Corresponding Author: Rehan Kapadia*.)

The authors are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, California 90089, United States (email: rkapadia@usc.edu).

effect transistors (FG-PFETs) and demonstrate that it can function as the computing element for an optical neural network (ONN), sensing and classifying images simultaneously. The floating gate is used to modulate the responsivity in the channel region, which mimics the behavior of a synapse. A simulated ONN indicated the accuracy up to ~94% for color-mix MNIST hand-written digits recognition. In addition, combined with the merits of non-volatility, back end of the line (BEOL) compatible process, and scalable material integration techniques, the FG-PFETs are demonstrated as potentially promising computing elements for machine vision.

II. Device Performance and Mechanism

A. Device Structure

The FG-PFETs were directly integrated on Si/SiO_2 wafer using a similar device architecture reported previously [28]. Here, the InP channel was chosen due to its direct bandgap property, good responsivity at visible wavelength, and high electron mobility. Moreover, the cost effective templated liquid phase (TLP) approach removed the requirement for single crystalline substrate, and enabled the monolithic InP integration directly on amorphous substrate even at low growth temperature[29, 30]. To be more specific, a patterned In/SiO_2 bilayer was deposited on the oxide substrate with photo lithography, thermal evaporation, and lift-off process. Then the samples were heated in a phosphine ambient, and supersaturated phosphorus in the liquid phase indium drives the a phase transformation from In to InP. By tuning the growth condition, it is possible to ensure that each channel region grows from a single nucleus. And with time, the entire area of is transformed to InP, retaining the same geometry as patterned. The fully grown InP channel is with dimension of 25 μm length, 6 μm width, and 300 nm thickness.

After the growth of InP channel, diluted hydrofluoric acid ($HF:H_2O$ = 1:10) was used to etch away the SiO_2 capping layer, and patterned 5 nm Ge, 20 nm Au, and 80 nm Ni were deposited by electron-beam evaporation as the source and drain contact. A 380℃ rapid thermal annealing in N_2 was carried out to reduce the contact resistance.

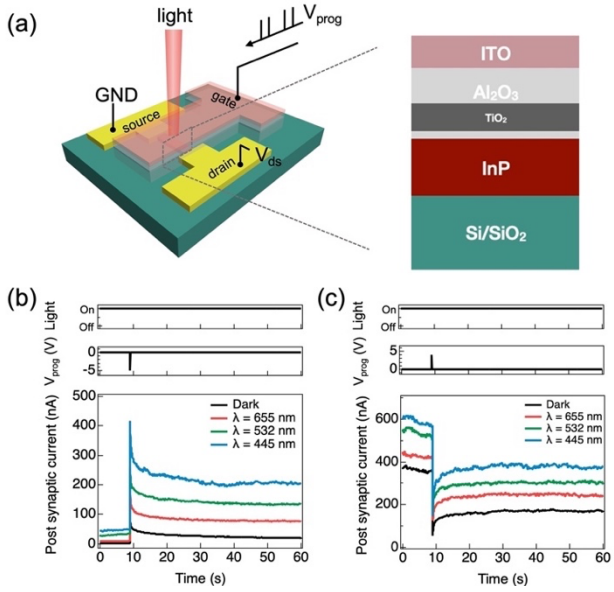

Fig. 1. Schematic and post-synaptic current of FG-PFETs. (a) Schematic of FG-PFETs and the cross-section film stack. (b) PSCs with one -5V V_{prog} pulse in dark, and illumination of three wavelengths ($\lambda = 655, 532, and\ 445\ nm$). (c) PSCs with one +4V V_{prog} pulse in dark, illumination of three wavelengths ($\lambda = 655, 532, and\ 445\ nm$).

As shown by the schematic in Fig. 1(a), the heterogeneous structured dielectric stack composed of 5 nm Al_2O_3, 5 nm TiO_2, and 50 nm Al_2O_3 were deposited by atomic layer deposition (ALD) at 200℃, it enabled the devices with both static and time dependent programmable conductance in a wide temporal scale of 10^{-3} to 10^5s. This dielectric stack is similar with Oxide-Nitride-Oxide (ONO) structure used in industry[31]. And the indium tin oxide (ITO) was deposited as the gate electrode to enable the interaction between InP channels and optical stimulations, due to its wide bandgap (3.5-4.3 eV) and the high transmission in the visible light spectrum. The representative output and transfer curves are shown in Methods. Notably, this device structure can work as the computing element for spiking neural networks as well, even without interference with optical spikes[28].

B. Device Performance

The synaptic behavior under electrical and optical stimulation was investigated by exposing the gate terminal to programming voltage pulses (V_{prog}), while maintaining the source drain bias, V_{ds} = 0.4V, and illuminating the device with wavelengths of 655, 532, and 445 nm. The post-synaptic currents (PSCs) are measured when light is incident on the channel area. The synaptic potentiation and depression of the device under illuminations are shown in Fig. 1(b) and Fig. 1(c). For this measurement, the optical signal with 0.3 mW/cm^2 power density was kept on during the measurement, and a single 10 ms long -5V/+4V V_{prog} pulse was applied. Independent of the illumination condition, the PSCs showed modulated conductance of the channels up to 50s after the V_{prog} pulse, suggesting the number of trapped or detrapped carriers in the heterogeneous dielectric structure sustained more than 50s. Notably, the segregated PSCs resulting from the wavelength of irradiations is expected due to the wavelength dependent absorption coefficient for the thin InP channels.

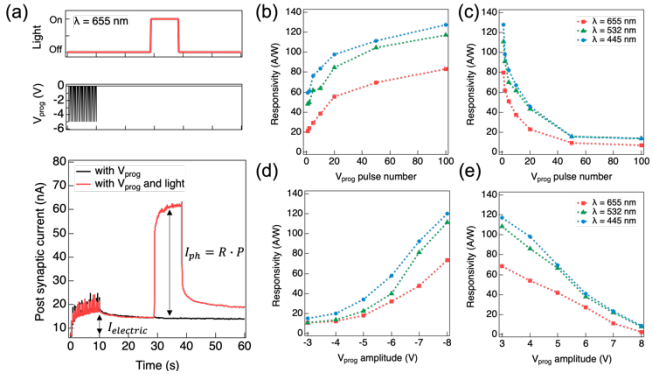

Fig. 2. Programmable responsivities of FG-PFETs. (a) PSCs with 20 pieces of -5V V_{prog} pulse and ~10s light pulse ($\lambda = 655\ nm$) 20s after the V_{prog}. (b)-(e) The relationship of responsivities of FG-PFETs extracted from (b) increased number of potentiation V_{prog} pulses, (c) increased number of depression V_{prog} pulses, (d) decreased amplitude of V_{prog} in the negative region, and (e) increased amplitude of V_{prog} in the positive region.

Unlike the floating-gate photo synapses that rely on the injection of carriers into the floating-gate via optical pulses [26, 27], the responsivity of FG-PFETs can be directly programmed by electrical spikes, which potentially enables fast and energy efficient image recognition. To demonstrate this property, electrical and optical stimuli are split in temporal space as shown in the upper panel of Fig. 2(a). Twenty -5V V_{prog} pulses were applied to the gate of the FG-PFETs and followed by a 10s optical pulse with 20s interval time. As indicated by the bottom graph in Fig. 2(a). The FG-PFETs demonstrated good overall responsivity of 55.6 A/W. The responsivity was calculated by measuring the PSC under the dark and light and then subtracting the values. The relationships between responsivity and V_{prog} pulse number and amplitude are extracted and shown in Fig. 2(b) to (e). Clearly, the responsivities of the FG-PFETs can be tuned up by the increased number of -5V V_{prog} pulses, and can be tuned down with +5V amplitude. Here we argue that the charge trapping in the dielectric gate modifies the transconductance of the channel at a fixed gate bias, thus modifying the photogating effect. Additionally, photoexcitation of the channel provides additional carries which may be trapped in the dielectric/at the interface. Therefore, the responsivities demonstrated saturation behavior in Fig. 2(b) and (c) with increased pulse number. It is observed that the responsivity changes more dramatically when the pulse number is small (<20), and then begins to saturate when the pulse number is larger than 20. The behavior versus programming pulse voltage does not show any saturation, as shown in Fig. 2(d) and (e), as the increase V_{prog} amplitude enables a greater density of carriers to be trapped or detrapped from the charge-trapping dielectric. For each measurement, pre-conditioning electrical stimuli were applied to reset the FG-PFEs to ground states (see Methods). Here, the wavelength-related segregation was also observed and generally matches the trends observed in Fig. 1(b) and (c).

C. *Device Mechanism*

The tunable responsivity in FG-PFETs is attributed to the combination of charger trapping in the heterostructured gate dielectric and the photogating mechanism of the InP channel region. As demonstrated in previous work [28], a TiO_2 layer in an Al_2O_3 "cladding layer" provides a potential well, and allows long-term carrier trapping. The total trapped charge in the TiO_2 layer is modified by the V_{prog} pulses applied to the gate, modifying the threshold voltage as shown in Fig. 3(a).

Photoconductive gain is widely observed in the low-dimensional semiconductor-based photoconductors [32]. The equation $G = {\tau}/{\tau_t}$ (τ is the minority carrier lifetime, τ_t is the carrier transit time) is commonly used to estimate the gain. Another way is using $G = (I_{PH}/e)/(PA \cdot \eta/h\nu)$, where P is the incident power density, $h\nu$ is the incident photon energy, A is the irradiated area, and η is the quantum efficiency defined as the product of light absorption efficiency and charge transfer efficiency [32-34]. Taking the measured results from the FG-PFETs shown in Fig. 2(a) and assumed $\eta \sim 45.1\%$ (see Methods) into the equation resulted in a photoconductive gain $G \sim 230$. Conventionally, photoconductive gain originated from the distinct mobility of electrons and holes. The net photocurrent $\Delta I = \Delta\sigma A\frac{V}{l} = q(\Delta n \cdot \mu_n + \Delta p \cdot \mu_p)A\frac{V}{l}$, where Δn and Δp are the excess electron and hole concentrations, μ_n and μ_p are the electron and hole mobilities, A is the semiconductor channel cross-section area, l is the channel length, and V is the bias provided. Consider the low-level injection, then the excess electron or hole ($\Delta n = \Delta p$) concentration can be described by ${\partial \Delta n}/{\partial t} = g(t) - {\Delta n}/{\tau_n}$, where $g(t)$ is the net optical generation rate, and τ_n is the excess electron lifetime. If $g(t)$ is independent of time, and at ${\partial \Delta n}/{\partial t} = 0$ state, then $\Delta n = g\tau_n$. With additional assumption: $\mu_n \gg \mu_p$, and transit time of electrons $\tau_t = \frac{l}{v_n} = \frac{l^2}{\mu_n V}$, it yields: $\Delta I \cong q\mu_n g\tau_n A\frac{V}{l} = qglA\frac{\tau_n}{\tau_t}$, and the term $\frac{\tau_n}{\tau_t}$ is in the same photoconductive gain expression for G. It can be understood that with incident light, electron-hole pairs are generated and separated by the applied bias. A fraction of the electrons and holes recombine immediately, and a fraction move in the opposite direction due to the electric field. However, if the transit time for an electron is shorter than the excess minority carrier lifetime, when excess electrons reach the anode, the same amount of electrons enter the photoconductor from the cathode to maintain the charge neutrality until the electrons recombine with the holes.

As for the photogating effect, although the equation of gain $G = {\tau}/{\tau_t}$ is identical, it describes the phenomenon that one type of the photogenerated carriers (holes in our case) is trapped in localized states, and work as a local gate voltage that influences the channel conductance. These states in the FG-PFETs may come from the inhomogeneous potential caused by defects and interfacial trap states. The relationship between the photocurrent and the local gate voltage can be written as: $\Delta I = \frac{\partial I_{ds}}{\partial V_g}\Delta V_g = g_m \Delta V_g$. For the FG-PFETs in this work, when plotting the measured gate voltage-dependent photocurrent and transconductance with the same range shown by Fig. 3(b), the photocurrent under 655nm illumination closely followed the transconductance trajectory during the V_g sweep from -2V to 6V. It suggests that photogating effect is the dominant mechanism in the devices.

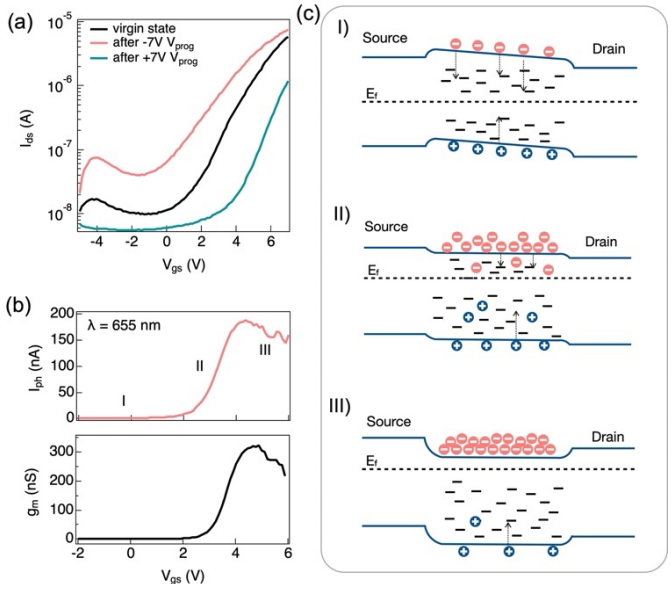

Fig. 3. The mechanisms behind the FG-PFETs. (a) I_{ds}-V_{gs} sweep before and after applying 20 +7V and -7V V_{prog} pulses. (b) Dependency of I_{ph} to V_{gs} with 655 nm irradiation and transconductance swept at the same range. (c) Band diagram corresponding to the three working regions in (b).

For the better understanding, we can divide the gate voltage-dependent photocurrent into three regions, and the corresponding band diagram are demonstrated in Fig. 3(c). When the negative gate voltage is applied to the device, Fermi level is close to mid-gap. The trap states above the Fermi level can capture electrons, and the states with energies below the Fermi level are able to capture holes. At this condition, the numbers of trap states for electrons and holes are close to each other, which results in a similar number of both kinds of carriers and small photocurrent (region I in Fig. 3(c)). With the increased V_g, the Fermi level move closer to the conduction band causes the increase of the electrons in the channel region, therefore, more electron trap states are occupied. On the other side, the decreased number of holes causes most hole traps to be accessible. In this condition, more photogenerated holes are trapped than photogenerated electrons and it causes the strong photogating effect manifested by the higher photocurrent with increased V_g as show in region II in Fig. 3(c). However, when the V_g increases to a higher value (region III), due to the finite number of hole traps and decreased carrier transport efficiency because of the higher metal-semiconductor barrier, photocurrent reaches its maximum point and starts to drop. In short, due to the modulation of the number of available trap states for carriers and the carrier transport efficiency, the responsivity of the device demonstrates strong dependency of V_g. This also explained why the V_{prog} pulse number and amplitude can program the responsivity with the heterogeneous gate dielectrics structure in the devices.

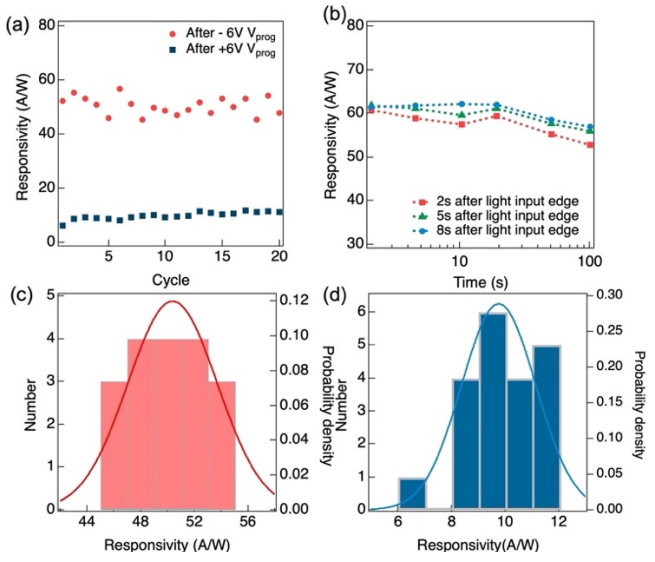

Fig. 4. Variation and duration of programmed responsivity. (a) Responsivities of FG-PFETs when circulating +6V V_{prog} pulse train and -6V V_{prog} pulse train. (b) Responsivities of FG-PFETs extracted from 2, 5, and 8s after illuminating the light pulse, and their dependencies to interval time between V_{prog} and irradiation. (c) and (d) Histogram of HRS and LRS, and their Gaussian distribution fitting curve.

D. Statistical Performance

The variation for programming the responsivity of the FG-PFETs was examined by circulating positive and negative V_{prog} pulses shown in Fig. 4(a). Twenty +6V V_{prog} pulses were applied first to the gate electrode with 0.4V bias between the source and the drain, and 655 nm laser with power density 0.3 mW/cm^2 was irradiated to the channel region 20s after the last piece of V_{prog} pulse. The duration of the illumination is 10s. After 30s, twenty -6V V_{prog} pulses were applied to the gate with the same V_{ds} and illumination. This measurement was looped 20 times and the results are shown in Fig. 4(a). Two distinct responsivity levels, high responsivity state (HRS) around 50 A/W, and low responsivity state (LRS) around 9.8 A/W are demonstrated. The variations for both states are low, and their distribution are fit using a Gaussian distribution as illustrated in Fig. 4(c) and (d). This results in the mean (μ) and standard deviation (σ) to be $\mu_{HRS} = 50.39$, $\sigma_{HRS} = 3.33$ and $\mu_{LRS} = 9.83$, $\sigma_{HRS} = 1.39$, respectively. The values here were also used for estimate noise in the neural network simulation discussed in the later section.

To work as a programmable photosensor and carry out image processing offline, the programmed responsivities in the FG-PFETs need to persist till the input images are illuminated on the device array. In Fig. 4 (b), the responsivities are extracted when the interval time between the last piece of V_{prog} pulse and illumination was varied from 2 to 100s. The preconditioning and V_{prog} pulse train (20 pieces of +6 V V_{prog} pulses with 20s resting time, then followed by 20 pieces of -7V V_{prog} pulses) were applied to the devices, and $t_{interval}$ (= 2, 5, 10, 20, 50, 100s) after the last pulse, a 655 nm laser is irradiated for 10s. The responsivities were sampled at 2, 5, and 8s after the rising edge of the synaptic current caused by illumination. As shown in Fig. 4(b), the responsivities measured 100s after the V_{prog} slightly dropped to 87%-93% of their value measured at 2s after the V_{prog}. And the sampling timing only causes a 7% difference in the worst case (at $t_{interval}$= 10s).

III. Simulation of ONN

A. ONN Structure

Human beings can distinguish colorful objects or disentangle the overlayed colorful images without much effort. This task can be finished as well using sensor arrays built by InP-based FG-PFETs. The optical neural network (ONN) constructed by FG-PFETs was simulated as shown in Fig. 5. The MNIST dataset [35] was modified to generate input images for the ONN. As the examples in Fig. 5(a), 1500 handwritten number "7" and "1" images were picked out and stacked into red (R), green (G), and blue (B) channels, and generated 6 classes of input images (R7G1, R7B1, G7R1, G7B1, B7G1, and B7R1). The ONN was developed with the simple two-layer structure as indicated in Fig. 5(b). The input layer contained 784 neuron arrays with color-filtering function to mimic the biological cone cells of human vision system, this step can be implemented by directly illuminating the colorful images to the 28×28 pixel array consists of $28 \times 28 \times 6$ FG-PFETs and color filters. As shown in Fig. 5(c), in each pixel, the input signal is separated by red, green, and blue filters, then sensed and processed by 6 FG-PFETs at the same pixel position. The synaptic connections

between the input layer and output layer are corresponding to the responsivity values of the 784× 6 FG-PFETs. Each output node is fully connected to one of the subpixels. Consequently, when the colorful image is directly irradiated to the FG-PFETs array, the sensing process and Kirchhoff's law conducted the matrix multiplication $P \cdot R = I$. After subtracting the dark current, the results can be processed by the SoftMax activation function: $\phi_m(I) = {e^{I_m \xi}}/{\sum_{k=1}^{M} e^{I_k \xi}}$ where ξ is a scaling factor, and cross-entropy $Loss = -\frac{1}{M}\sum_{m=1}^{M} y_m \log[\phi_m(I)]$ is used to calculate the loss from the output $\phi_m(I)$ and the label y_m.

The circuit diagram of the first three pixels is shown in Fig. 5(d). The other 781 pixels and corresponding output current that will merge with them are not shown here. At each pixel position, 6 FG-PFETs with 3 color filters sense and process the different wavelength signals and output 6 current values. Each current node measures the sum of 784 current values from each pixel due to parallel connection, then the current values will be processed by the SoftMax activation function in the software. Here the gate voltage V_{nm}, where n is pixel number, m is sub-pixel number, is 0 V for reading operation, and $V_{program}$ during the programming operation. This circuit is different with the crossbar array since the three-terminal devices are used, and the voltage values applied to the crossbar array typically carry the input information through parameters like amplitude, frequency, and duty cycles.

B. Update Rules

For each training and testing epoch, 1200 training images and 300 test images were shuffled, and the backpropagation and gradient descent approaches were implemented to know the tuning direction for the responsivity of each FG-PFET. The initial responsivities of the devices are randomly generated between the R_{max} and R_{min} with uniform distribution. We implemented the stepped updating rule based on the fitting curves of the measured responsivities dependency to V_{prog} pulse number [36]. More specifically the rules are:

When ${d(Loss)}/{dR} > 0$:

$$R_{n+1} = R_n + \Delta R_p = R_n + \alpha_p e^{-\beta_p \frac{R_n - R_{min}}{R_{max} - R_{min}}}$$

When ${d(Loss)}/{dR} = 0$:

$$R_{n+1} = R_n$$

When ${d(Loss)}/{dR} < 0$:

$$R_{n+1} = R_n + \Delta R_D = R_n - \alpha_D e^{-\beta_D \frac{R_{max} - R_n}{R_{max} - R_{min}}}$$

Where R_{n+1} is the target updating responsivity, R_n is the current responsivity, α_p and α_D are the fitting factors from the potentiation and depression curve, β_p and β_D are the non-linear factors, R_{max} and R_{min} are the turning range of the responsivities. When the R_{n+1} is out of the range between R_{max} and R_{min}, clipping is applied during the value updating.

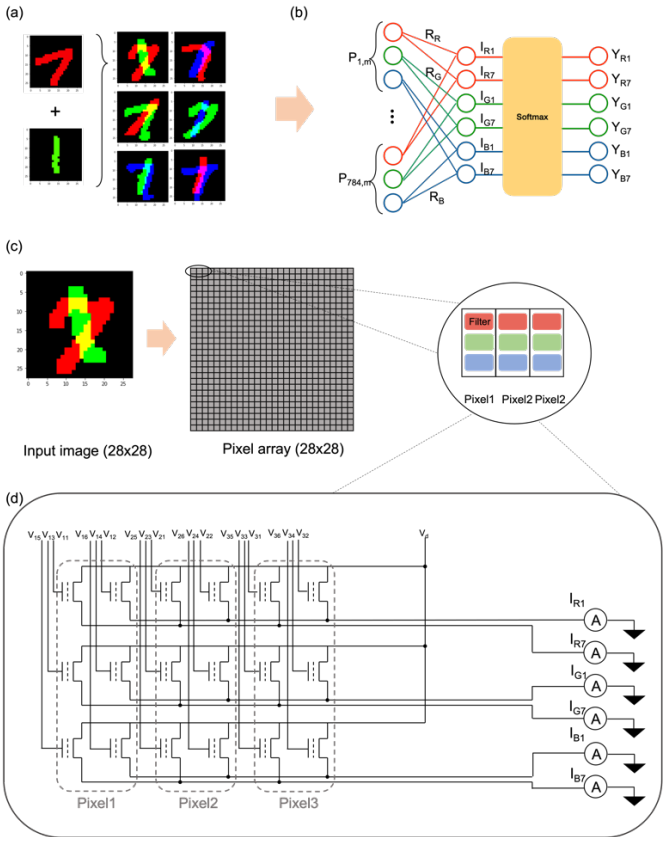

Fig. 5. ONN built with FG-PFETs. (a) Examples of stacked MNIST digits, working as the color-mixed input images to the ONN. (b) Schematic of the ONN built with FG-PFETs. (c) Schematic of pixel array, and pixel with color filters. (d) Circuit diagram of first 3 pixels, each pixel contain 6 FG-PFETs and three color filters.

C. Simulation Results

In addition, the updating noise based on the measured results and gaussian distribution fitting curve in Fig. 4(a) was included for the simulation. The resulted accuracy and loss over 40 training and testing epochs are shown in Fig. 6(c) and (d), plotted with projected results from the devices having lower updating noise. Clearly, within 10 training epochs, the accuracy can reach up to ~94%, higher than ONN reported in reference paper[36], and the loss drops rapidly even for the devices with the highest update noise. However, after 10 epochs, lower update noise led to higher accuracy and more stable performance. This is due to the update noise contributing more to the overall number of available states and effect of small weight updates [37]. Overall, the simulation results based on the measured performance of FG-PFETs suggest that the ONN constructed by FG-PFETs can be developed for color-mixed handwritten digits recognition, and the accuracy can be further improved by lowering the responsivity updating noise.

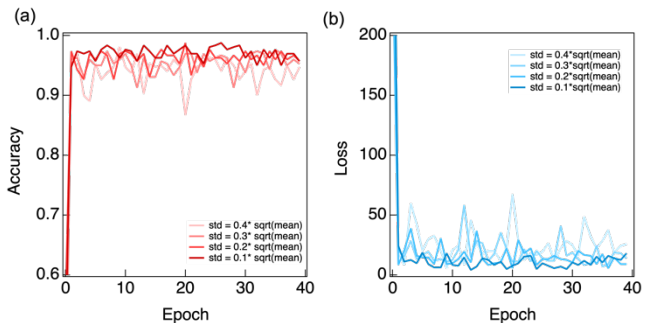

Fig.6 (a) Accuracy and (b) loss of the simulated ONN over 40 epochs of training and testing.

IV. Conclusion

In this work, top-gated FG-PFETs based on single crystalline InP channels grown by the TLP approach were reported. A floating gate is shown to modify the responsivity of the device by tuning the photogating effect, and enabling programming of the responsivities of the FG-PFETs. It was shown by both the number or amplitude of V_{prog} pulses could be used for this, but it is expected that pulse number programming is more practical for real circuits. Using this behavior, the FG-PFETs was shown to perform the sensing and classification simultaneously when working as the computing element in an ONN. Due to the large tuning range, and low variation of FG-PFETs, the constructed ONN shows a steep decline of loss and the image recognition accuracy ramp-up to ~94% during training and testing epochs. When coupled with the BEOL compatible LT-TLP approach, InP based FG-PFETs may pave the way to more energy efficient hardware level machine vision.

V. Methods

A. Output and Transfer Curves

Fig. S1 below shows the representative output curves and transfer curves of PF-PFETs in this work.

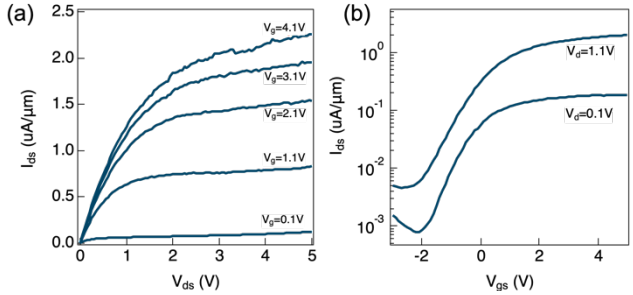

Fig. S1. (a) Output curves of FG-PFETs with V_g from 0.1 to 4.1V. (b) Transfer curves of FG-PFETs with V_d=0.1 and 1.1V.

B. Ground States Alignment

To align the device status before applying various V_{prog} pulse train, pre-conditioning pulses are applied. When the V_{prog} pulse trains use the negative amplitude like for Fig. 2(b) and (d), 20 pieces of +7V pulses with 10ms pulse width, 1s pulse period, and 20sec resting time are applied. And when the amplitude is positive as for the measurements for Fig. 2 (c) and (e), 20 pieces of -7V pulses with 10ms pulse width, 1s pulse period, and 20sec resting time are applied. All of them used the V_{ds}=0.4V.

C. Quantum Efficiency estimation

η was calculated to be 45.1% using the Lambert-Beer's law : $A(\omega) = 1 - e^{-\alpha(\omega)\Delta Z}$ where $A(\omega)$ is the material absorption, ω is the light frequency, ΔZ = 300 nm is the thickness of the InP channels. Here the charge transfer efficiency is assumed to be 1.

REFERENCES

[1] L. Mennel, J. Symonowicz, S. Wachter, D. K. Polyushkin, A. J. Molina-Mendoza, and T. Mueller, "Ultrafast machine vision with 2D material neural network image sensors," *Nature,* vol. 579, no. 7797, pp. 62-66, Mar, 2020.

[2] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature,* vol. 575, no. 7784, pp. 607-617, Nov, 2019.

[3] 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, "True North: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *Ieee Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 34, no. 10, pp. 1537-1557, Oct, 2015.

[4] F. C. Zhou, Z. Zhou, J. W. Chen, T. H. Choy, J. L. Wang, N. Zhang, Z. Y. Lin, S. M. Yu, J. F. Kang, H. S. P. Wong, and Y. Chai, "Optoelectronic resistive random access memory for neuromorphic vision sensors," *Nature Nanotechnology,* vol. 14, no. 8, pp. 776-782, Aug, 2019.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, May, 2015.

[6] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the Ieee,* vol. 86, no. 11, pp. 2278-2324, Nov, 1998.

[7] K. Berggren, Q. F. Xia, K. K. Likharev, D. B. Strukov, H. Jiang, T. Mikolajick, D. Querlioz, M. Salinga, J. R. Erickson, S. Pi, F. Xiong, P. Lin, C. Li, Y. Chen, S. S. Xiong, B. D. Hoskins, M. W. Daniels, A. Madhavan, J. A. Liddle, J. J. McClelland, Y. C. Yang, J. Rupp, S. S. Nonnenmann, K. T. Cheng, N. A. Gong, M. A. Lastras-Montano, A. A. Talin, A. Salleo, B. J. Shastri, T. F. de Lima, P. Prucnal, A. N. Tait, Y. C. Shen, H. Y. Meng, C. Roques-Carmes, Z. G. Cheng, H. Bhaskaran, D. Jariwala, H. Wang, J. M. Shainline, K. Segall, J. J. Yang, K. Roy, S. Datta, and A. Raychowdhury, "Roadmap on emerging hardware and technology for machine learning," *Nanotechnology,* vol. 32, no. 1, Jan, 2021.

[8] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element," *Ieee Transactions on Electron Devices,* vol. 62, no. 11, pp. 3498-3507, Nov, 2015.

[9] H. S. P. Wong, H. Y. Lee, S. M. Yu, Y. S. Chen, Y. Wu, P. S. Chen, B. Lee, F. T. Chen, and M. J. Tsai, "Metal-Oxide RRAM," *Proceedings of the Ieee,* vol. 100, no. 6, pp. 1951-1970, Jun, 2012.

[10] J. J. S. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," *Nature Nanotechnology,* vol. 8, no. 1, pp. 13-24, Jan, 2013.

[11] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H. S. P. Wong, "Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing," *Nano Letters,* vol. 12, no. 5, pp. 2179-2186, May, 2012.

[12] S. Dutta, C. Schafer, J. Gomez, K. Ni, S. Joshi, and S. Datta, "Supervised Learning in All FeFET-Based Spiking Neural Network: Opportunities and Challenges," *Frontiers in Neuroscience,* vol. 14, pp. 14, Jun, 2020.

[13] G. D. Feng, J. Jiang, Y. R. Li, D. D. Xie, B. B. Tian, and Q. Wan, "Flexible Vertical Photogating Transistor Network with an Ultrashort Channel for In-Sensor Visual Nociceptor," *Advanced Functional Materials,* vol. 31, no. 36, Sep, 2021.

[14] G. D. Feng, J. Jiang, Y. H. Zhao, S. T. Wang, B. Liu, K. Yin, D. M. Niu, X. H. Li, Y. Q. Chen, H. G. Duan, J. L. Yang, J. He, Y. L. Gao, and Q. Wan, "A Sub-10 nm Vertical Organic/Inorganic Hybrid Transistor for Pain-Perceptual and Sensitization-Regulated Nociceptor Emulation," *Advanced Materials,* vol. 32, no. 6, Feb, 2020.

[15] Y. C. Cheng, H. J. W. Li, B. Liu, L. Y. Jiang, M. Liu, H. Huang, J. L. Yang, J. He, and J. Jiang, "Vertical 0D-Perovskite/2D-MoS(2)van der Waals Heterojunction Phototransistor for Emulating Photoelectric-Synergistically Classical Pavlovian Conditioning and Neural Coding Dynamics," *Small,* vol. 16, no. 45, Nov, 2020.

[16] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. N. Piramanayagam, "Spintronics based random access memory: a review," *Materials Today,* vol. 20, no. 9, pp. 530-548, Nov, 2017.

[17] D. Sarkar, J. Tao, W. Wang, Q. F. Lin, M. Yeung, C. H. Ren, and R. Kapadia, "Mimicking Biological Synaptic Functionality with an Indium Phosphide Synaptic Device on Silicon for Scalable Neuromorphic Computing," *Acs Nano,* vol. 12, no. 2, pp. 1656-1663, Feb, 2018.

[18] T. Chang, S. H. Jo, and W. Lu, "Short-Term Memory to Long-Term Memory Transition in a Nanoscale Memristor," *Acs Nano,* vol. 5, no. 9, pp. 7669-7676, Sep, 2011.

[19] A. J. Arnold, A. Razavieh, J. R. Nasr, D. S. Schulman, C. M. Eichfeld, and S. Das, "Mimicking Neurotransmitter Release in Chemical Synapses via Hysteresis Engineering in MoS2 Transistors," *Acs Nano,* vol. 11, no. 3, pp. 3110-3118, Mar, 2017.

[20] Q. F. Xia, and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing," *Nature Materials,* vol. 18, no. 4, pp. 309-323, Apr, 2019.

[21] P. Yao, H. Q. Wu, B. Gao, J. S. Tang, Q. T. Zhang, W. Q. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature,* vol. 577, no. 7792, pp. 641-646, Jan, 2020.

[22] P. M. Sheridan, F. X. Cai, C. Du, W. Ma, Z. Y. Zhang, and W. D. Lu, "Sparse coding with memristor networks," *Nature Nanotechnology,* vol. 12, no. 8, pp. 784-789, Aug, 2017.

[23] C. Li, M. Hu, Y. N. Li, H. Jiang, N. Ge, E. Montgomery, J. M. Zhang, W. H. Song, N. Davila, C. E. Graves, Z. Y. Li, J. P. Strachan, P. Lin, Z. R. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. F. Xia, "Analogue signal and image processing with large memristor crossbars," *Nature Electronics,* vol. 1, no. 1, pp. 52-59, Jan, 2018.

[24] X. M. Zhang, Y. Zhuo, Q. Luo, Z. H. Wu, R. Midya, Z. R. Wang, W. H. Song, R. Wang, N. K. Upadhyay, Y. L. Fang, F. Kiani, M. Y. Rao, Y. Yang, Q. F. Xia, Q. Liu, M. Liu, and J. J. Yang, "An artificial spiking afferent nerve based on Mott memristors for neurorobotics," *Nature Communications,* vol. 11, no. 1, pp. 9, Jan, 2020.

[25] T. D. Fu, X. M. Liu, H. Y. Gao, J. E. Ward, X. R. Liu, B. Yin, Z. R. Wang, Y. Zhuo, D. J. F. Walker, J. J. Yang, J. H. Chen, D. R. Lovley, and J. Yao, "Bioinspired bio-voltage memristors," *Nature Communications,* vol. 11, no. 1, Apr, 2020.

[26] H. L. Park, H. Kim, D. Lim, H. Zhou, Y. H. Kim, Y. Lee, S. Park, and T. W. Lee, "Retina-Inspired Carbon Nitride-Based Photonic Synapses for Selective Detection of UV Light," *Advanced Materials,* vol. 32, no. 11, Mar, 2020.

[27] Y. Wang, Z. Y. Lv, J. R. Chen, Z. P. Wang, Y. Zhou, L. Zhou, X. L. Chen, and S. T. Han, "Photonic Synapses Based on Inorganic Perovskite Quantum Dots for Neuromorphic Computing," *Advanced Materials,* vol. 30, no. 38, Sep, 2018.

[28] J. Tao, D. Sarkar, S. Kale, P. K. Singh, and R. Kapadia, "Engineering Complex Synaptic Behaviors in a Single Device: Emulating Consolidation of Short-term Memory to Long-term Memory in Artificial Synapses via Dielectric Band Engineering," *Nano Letters,* vol. 20, no. 10, pp. 7793-7801, Oct, 2020.

[29] D. Sarkar, S. Z. Weng, D. Z. Yang, J. Tao, S. Naik, S. Kale, H. U. Chae, Y. P. Xu, B. Mesri, S. B. Cronin, F. Greer, and R. Kapadia, "Low Temperature Growth of Crystalline Semiconductors on Nonepitaxial Substrates," *Advanced Materials Interfaces,* vol. 7, no. 14, pp. 5, Jul, 2020.

[30] J. Tao, D. Sarkar, S. Z. Weng, T. Orvis, R. Ahsan, S. Kale, Y. P. Xu, H. U. Chae, F. Greer, J. Ravichandran, C. Sideris, and R. Kapadia, "High mobility large area single crystal III-V thin film templates directly grown on amorphous SiO2 on silicon," *Applied Physics Letters,* vol. 117, no. 4, Jul, 2020.

[31] M. Joodaki, "Uprising nano memories: Latest advances in monolithic three dimensional (3D) integrated Flash memories," *Microelectronic Engineering,* vol. 164, pp. 75-87, Oct, 2016.

[32] H. H. Fang, and W. D. Hu, "Photogating in Low Dimensional Photodetectors," *Advanced Science,* vol. 4, no. 12, Dec, 2017.

[33] Q. S. Guo, A. Pospischil, M. Bhuiyan, H. Jiang, H. Tian, D. Farmer, B. C. Deng, C. Li, S. J. Han, H. Wang, Q. F. Xia, T. P. Ma, T. Mueller, and F. N. Xia, "Black Phosphorus Mid-Infrared Photodetectors with High Gain," *Nano Letters,* vol. 16, no. 7, pp. 4648-4655, Jul, 2016.

[34] S. L. Chuang, *Physics of photonic devices*: John Wiley & Sons, 2012.

[35] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 141-142, 2012.

[36] S. Seo, S. H. Jo, S. Kim, J. Shim, S. Oh, J. H. Kim, K. Heo, J. W. Choi, C. Choi, D. Kuzum, H. S. P. Wong, and J. H. Park, "Artificial optic-neural synapse for colored and color-mixed pattern recognition," *Nature Communications,* vol. 9, Nov, 2018.

[37] S. Agarwal, S. J. Plimpton, D. R. Hughart, A. H. Hsia, I. Richter, J. A. Cox, C. D. James, M. J. Marinella, and Ieee, "Resistive Memory Device Requirements for a Neural Algorithm Accelerator," *IEEE International Joint Conference on Neural Networks (IJCNN).* pp. 929-938, 2016.

Jun Tao is currently a Ph.D. candidate under the supervision of Prof. Rehan Kapadia in the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California. He received his B.S degree in Electronic Science and Technology and M.S degree in Microelectronics and Solid Electronics from Shanghai University, China in 2013 and 2016 respectively. He is interested in advanced materials and fabrication techniques for nanostructured photonic and electronic devices.

Juan Sanchez Vazquez is currently a Ph.D student under the supervision of Prof. Rehan Kapadia in the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California. He received his B.Sc. degree in Electrical Engineering from the University of Southern California in 2021. His research interests lie in the fabrication of semiconductor materials and devices, and their application to emerging hardware for the acceleration of AI algorithms.

Hyun Uk Chae is currently a Ph.D student under the supervision of Prof. Rehan Kapadia in the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California. He received his B.S degree in Electrical and Electronics Engineering from Dongguk University, South Korea in 2017. He is interested in fabrication and material growth techniques for electronic, photonic, and electrochemical devices.

Ragib Ahsan is a PhD student working in Professor Kapadia's group. He graduated with his BS degree in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology in 2017. His research interest lies in understanding the physics of semiconductor devices and implementing them experimentally, with a focus on hot electron devices.

Rehan Kapadia joined the faculty of the University of Southern California in the Ming Hsieh Department of Electrical and Computer Engineering in July 2014 where he presently holds the Colleen and Roberto Padovani Early Career Chair. He received his bachelors in electrical engineering from the University of Texas at Austin in 2008, and his Ph.D. in electrical engineering from the University of California, Berkeley in 2013. During his time at Berkeley, he was a National Science Foundation Graduate Research Fellow and winner of the David J. Sakrison Memorial Prize for outstanding research. Since joining USC he has received an Air Force Young Investigator Program award and an Office of Naval Research Young Investigator Program award. He has also been awarded an American Vacuum Society Peter Mark Memorial Award for young scientists. His interests lie at the intersection of material science and electrical engineering focusing on non-equilibrium electron devices, and materials growth technologies for next-generation electronic and photonic devices.