

Ultra-low power in-sensor neuronal computing with oscillatory retinal neurons for frequency-multiplexed, parallel machine vision

Ragib Ahsan

University of Southern California <https://orcid.org/0000-0002-3833-7851>

Hyun Uk Chae

University of Southern California

Seyedeh Atiyeh Abbasi Jalal

University of Southern California

Jun Tao

University of Southern California <https://orcid.org/0000-0003-2156-1731>

Subrata Das

University of Southern California

Hefei Liu

University of Southern California <https://orcid.org/0000-0001-6533-7112>

Jiang-Bin Wu

University of Southern California <https://orcid.org/0000-0002-8751-7082>

Stephen Cronin

University of Southern California

Han Wang

University of Southern California

Constantine Sideris

University of Southern California

Rehan Kapadia (✉ rkapadia@usc.edu)



rkapadia@usc.edu <https://orcid.org/0000-0002-7611-0551>

Physical Sciences - Article

Keywords:

Posted Date: June 14th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-2935296/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

single hidden layer neural network, approximating a liquid state machine. This network demonstrated a 14.8% reduction in classification error from 2.16% to 1.84% compared to the same neural network with standard photodetector inputs. Finally, the equivalent energy consumption to carry out image processing operations, including peripherals such as the Fourier transform circuits, is projected to be ~ 24 aJ/OP.

Main

In-sensor computing has emerged as a promising approach to improve computational speed and reduce energy consumption¹⁻¹⁵. Local weighting devices or tunable responsivity sensors enable in-sensor architectures, where the input signal is multiplied by a weight at the point of sensing, resulting in local multiply-accumulate (MAC) operations on the inputs. By eliminating the initial data conversion, storage, and transmission, in-sensor architectures offer dramatically higher speed and lower power consumption when compared to traditional von Neumann architectures. A wide variety of modalities, including auditory^{10,17-22}, olfactory¹⁹, tactile²⁴⁻²⁶ and vision^{6,7,9,11-13,15,27} sensors, benefit from the improved performance. However, these approaches generally execute a single MAC operation on the input data^{1,14,15,28-30}. Furthermore, parallel operations require scaling the number of weighting devices connected to each sensor, which can be costly from an area and power perspective.

We introduce an in-sensor computing approach where a coupled photosensor array carries out parallel computation on the input image. Each pixel in the array acts as an oscillator, generating an optical power-dependent frequency spectrum. When coupled, neighboring pixels also affect each pixel's frequency spectrum. The power in a frequency band then becomes a non-linear

function of the inputs. Separate frequency bands, therefore, encode separate non-linear functions of the inputs in parallel. Here, each pixel is an oscillatory retinal neuron (ORN) that directly converts the input optical signal into voltage oscillations. We show through simulation and experiment that coupled ORN networks carry out approximations of both basic and advanced image processing functions, such as edge detection and image segmentation directly in the sensor, encoded by choice of frequency and bandwidth of the output filter. Notably, the ORNs do not require external electrical power, and when considering peripheral circuits such as buffers, selector circuits, and analog fast Fourier transform circuits, the *equivalent energy per operation can be as low as 24 aJ/OP*. Using the change in frequency spectrum instead of the phase dramatically relaxes the fabrication tolerance requirements compared to other approaches that rely on synchronization of oscillators, such as Ising machines³³⁻³⁹, leading to considerably greater scalability.

The ORNs are composed of two elements, (i) a photodetector that exhibits voltage-controlled negative differential resistance (NDR) under illumination and (ii) an inductive element that can drive an electrical oscillation by taking advantage of the instability of the NDR behavior. A semiconductor-graphene-metal (SGM) photodetector, schematically shown in Figure 1a, exhibits NDR in the detector's power generation regime. The device comprises a p-type silicon substrate, a Ti/Au (5 nm/100 nm) metal grid, and a graphene layer. Linear scale I-V measurements of a 1 mm × 1 mm device under dark and uniform optical illumination are shown in Figure 1b. In the dark, the device exhibits Schottky-diode behavior, while exhibiting NDR under illumination. Figure 1c shows the log-scale I-V curves, highlighting that the NDR is only observed under

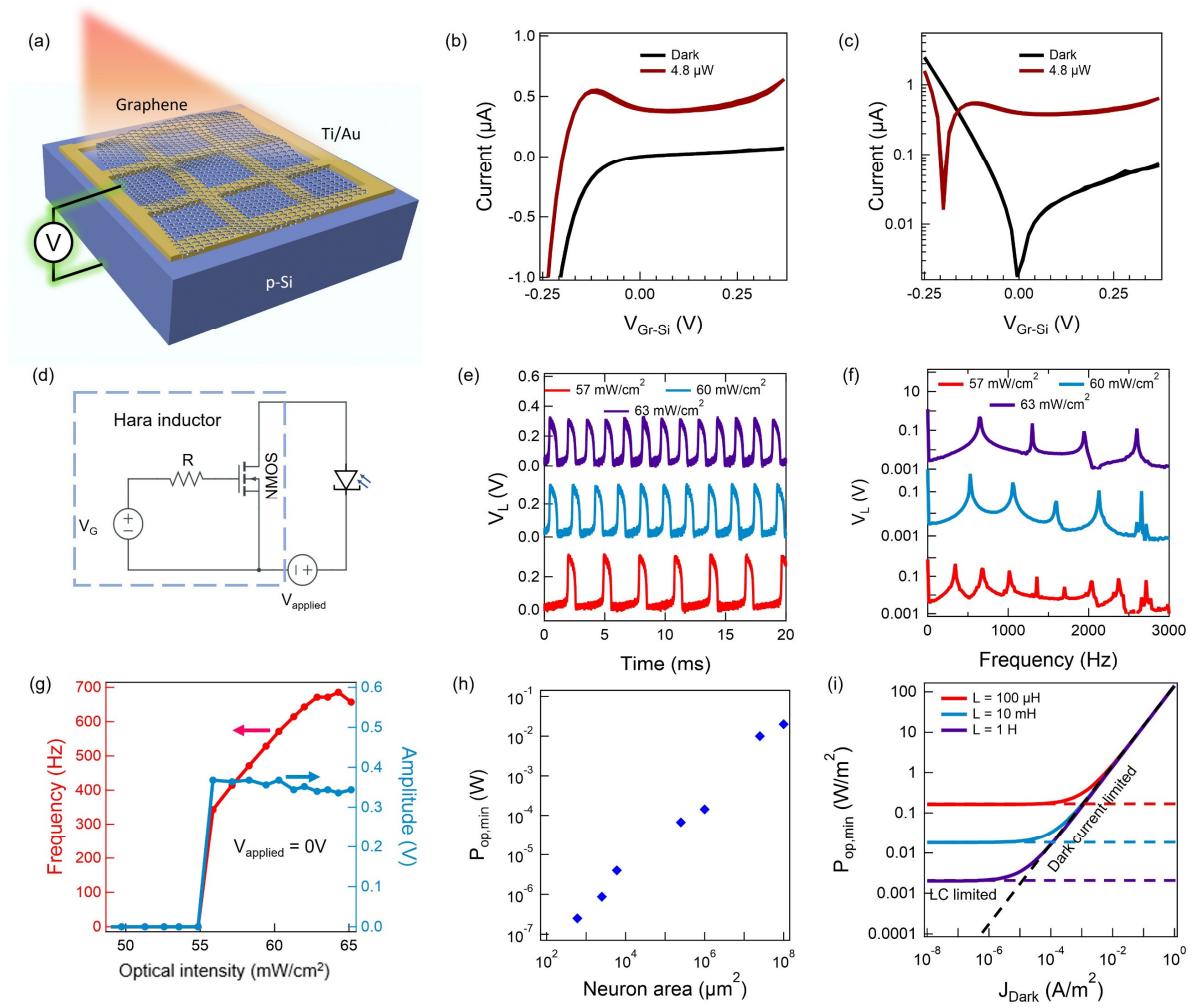


Figure 1: ORN enabled by SGM photodetector. (a) Schematic of the SGM photodetector device. (b) I-V curves measured at dark conditions and under uniform illumination (445 nm) in linear and (c) log scale. (d) Schematic of a single unit of ORN. (e) V-t curves measured at different optical intensities and (f) corresponding frequency spectrum. (g) spiking frequency and amplitude as a function of optical intensity. (h) Experimental plot of minimum optical power required for oscillation with neuron area. (i) Calculation of dark current limited and LC limited $P_{op,min}$ for oscillation without external electrical power.

illumination. Supplementary section S1 and supplementary Figures S1-7 discuss the device-level behaviors in detail. Connecting this device with an inductive element under appropriate bias conditions generates optical intensity dependent oscillations, as shown schematically in Figure 1d. An active inductive element, the Hara inductor, comprising a single MOSFET and a resistor,

enables the scalability of the ORN. The observed oscillations are analogous to classical Van der Pol oscillators and the Fitzhugh-Nagumo model of neurons⁴¹⁻⁴⁴.

Other graphene-based photodetectors have exhibited NDR behavior, but all at a forward diode bias⁴⁵⁻⁵¹. However, this device generates an open-circuit voltage and exhibits NDR at negative and zero applied voltages. This critical distinction allows oscillations at $V_{\text{applied}} \leq 0\text{V}$, which enables operation without external electrical power. Figure 1e shows experimental V-t curves for a photodetector with an active area of 1 cm^2 . Figure 1f shows the corresponding frequency spectra, illustrating the change as a function of the optical intensity. Figure 1g shows the oscillation frequency and amplitude as a function of incident optical intensity, where we observe that a minimum optical intensity is required to trigger oscillations in this ORN circuit. These measurements were all performed at $V_{\text{applied}} = 0\text{V}$.

To explore the scaling behavior of ORNs, photodetectors with areas between $600\text{ }\mu\text{m}^2$ and 1 cm^2 have been fabricated and tested. The minimum optical power required for oscillation *without external electrical power* scales linearly with the device area, as shown in Figure 1h. Two parameters limit the oscillation dynamics of ORNs, the dark current and the capacitance. First, the dark current does not exhibit NDR and adds with the light current. Second, the photon flux should generate sufficient light current so that the valley of the NDR is greater than the dark current. There must also be sufficient photocurrent to charge and discharge the capacitance at timescales of the oscillation frequency. The addition of external power can mitigate this limitation. For a moderately doped p-Si substrate, the depletion capacitance at the graphene-silicon junction is $\sim 0.1\text{ fF}/\mu\text{m}^2$. Figure 1i shows the minimum optical intensity for oscillation assuming a device capacitance of $0.1\text{ fF}/\mu\text{m}^2$ as a function of device dark current density. We can see a crossover between two different regimes: (1) inductance-capacitance (LC) limited regime at smaller dark currents and (2)

dark current limited regime at larger dark currents. For our photodetectors, the Schottky nature of the junction results in a larger dark current, limiting the threshold optical intensity to $\sim 400 \text{ W/m}^2$. At smaller dark current densities, it is possible to decrease this threshold to below 2 mW/m^2 .

Next, we present a simple demonstration of how these coupled oscillators carry out computation. We use simulations of ORN circuits connected to bandpass filters to elucidate the behavior of coupled ORNs and how image processing occurs. We considered an ORN comprising a photodetector with an active area of 1 mm^2 connected to an external inductor ($L = 10 \text{ mH}$) with $V_{\text{applied}} = 0\text{V}$. We simulated the V-t curves of the ORN using the experimental photodetector capacitance and J-V values. The V-t output of the simulation is filtered with varying center frequencies (f) and bandwidths (BW) representing different bandpass filters. Supplementary section S2 and Figure S8 show through simulation and analysis that each bandpass filtered output of a single ORN can be analytically approximated with Lorentzians. Figure 2a shows the schematic of two ORNs with inductive coupling, $L_C = 10 \text{ mH}$. Figure 2b plots the bandpass filtered V_{osc1} magnitude as a function of P_1 and P_2 for varying center frequencies $f = 28.4, 28, \text{ and } 27.6 \text{ KHz}$ with $\text{BW} = 200 \text{ Hz}$. The results show that two coupled oscillators define a curved subspace of the input. Figure 2c shows the simulation results for a fixed filter with $f = 28.4 \text{ KHz}$ and $\text{BW} = 200 \text{ Hz}$ and varying coupling impedance. This results in subspaces of varying shapes. While accurate solutions of the oscillator-coupled non-linear differential equations require numerical solutions, we can analytically approximate the subspace by reducing the two oscillator problem to a single oscillator problem by introducing a new quantity $P_{12} = \sqrt{P_1^2 + P_2^2 + a(P_1 + P_2) + kP_1P_2 + b}$ which nonlinearly combines P_1 and P_2 . The coupled oscillator result then becomes $V_{\text{osc}}(P_{12}, f, \text{BW}) = \frac{\gamma}{(P_{12} - P_{00})^2 + (\Delta P)^2}$, which can be fit to approximate the result from Figure 2c as

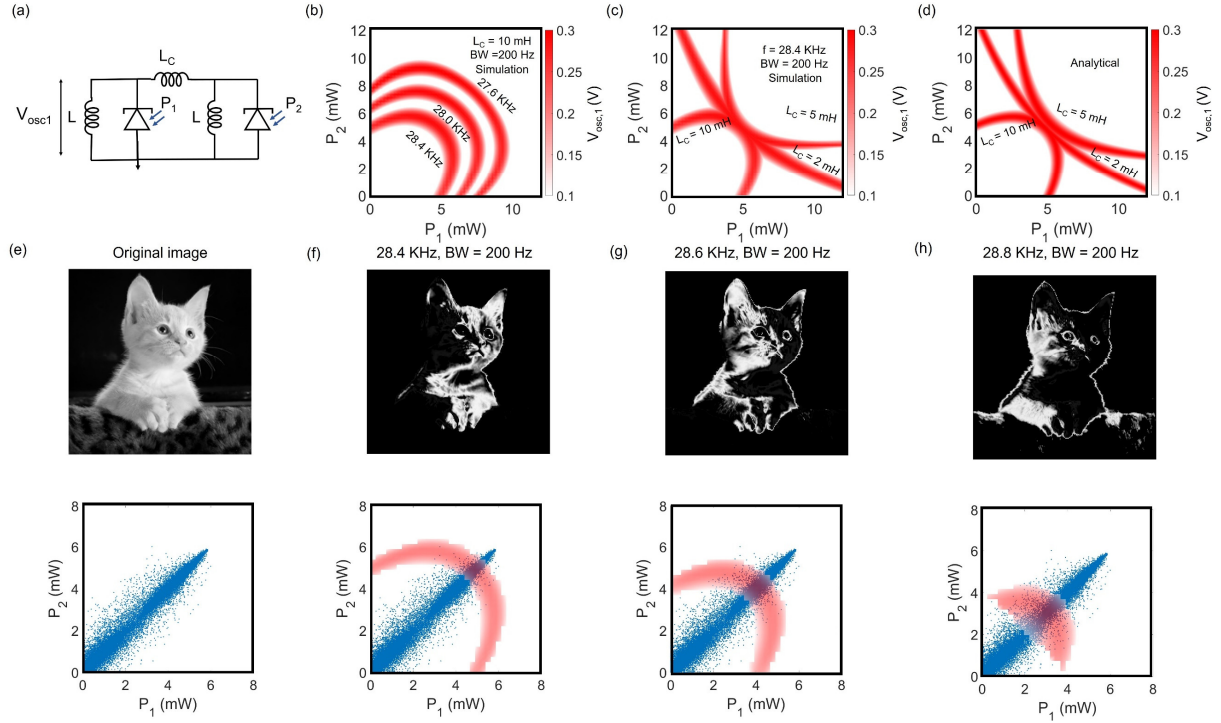


Figure 2: Frequency multiplexed computation with ORN. (a) Circuit schematic for two coupled ORNs. (b) ORN voltage colormap showing nonlinear peak surfaces and their shift at different center frequencies for $L_C = 10$ mH and $BW = 200$ Hz. (c) ORN voltage colormap showing different peak surface shapes for different L_C values and their (d) analytical approximations. (e) Original image and the scatter plot showing all the (P_1, P_2) pairs for this image if it were input to a 1×2 convolutional kernel. (f-h) Image transformations when the two coupled ORNs ($L_C = 10$ mH) receive the (P_1, P_2) pairs as inputs similar to a convolution operation and the corresponding scatter plots. The overlap between red and blue scatter plots show how different subsets of inputs are thresholded by the network at different center frequencies ($BW = 200$ Hz).

shown in Figure 2d. Here, P_{00} is a function of the center frequency f and ΔP is a function of the filter bandwidth, BW .

To obtain a visual representation of how an image is processed in this scheme, we have treated the 2-ORN circuit as a 1×2 convolutional kernel and processed a grayscale image of a cat (Fig. 2e, top panel) with 250×240 pixels. The bottom panel of Figure 2e shows the (P_1, P_2) pixel pairs, which serve as inputs to the 1×2 convolution kernel. The top panels in Figures 2f-h show the filtered output images for $f = 28.4, 28.6$ and 28.8 KHz and $BW = 200$ Hz. Clearly, the original image has been mapped to multiple processed images, indexed by the filter's center frequency. The

bottom panels of Figures 2f-h show how the subspaces, defined by the ORN coupling, filter center frequency (f), and bandwidth (BW), overlap with the (P_1 , P_2) pixel pairs of the original image. The coupled ORNs select the subset of the pixels that overlap with the defined subspace. These results on a toy problem visually show how non-linear computations are performed using coupled ORN oscillators.

To experimentally demonstrate how coupled ORNs carry out more useful and complex image processing functions, from edge detection to image sharpening, we have experimentally

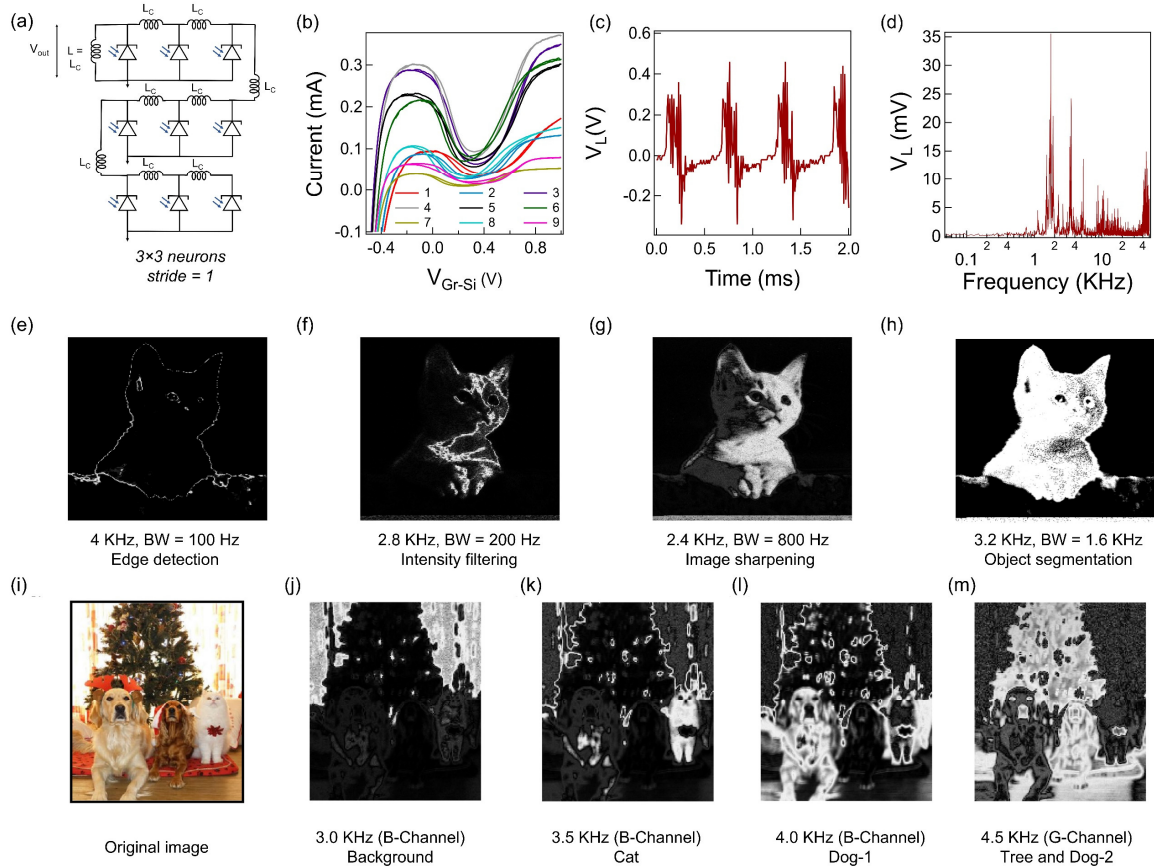


Figure 3: Image processing with coupled ORN network. (a) Circuit schematic for the ORN kernel (b) I-V curves of all 9 SGM detectors in the network under same optical illumination. (c) Oscillation V-t and (d) FFT curves at the output node when all ORNs are under uniform illumination. (e) Frequency band filtered images showing edge detection, (f) intensity filtering, (g) image sharpening, (h) object segmentation. (i) Original color image and frequency domain images showing (j-m) image segmentation operation.

fabricated a 3×3 ORN focal plane array with a cascaded connection, as shown in Figure 3a. We use this as a kernel that slides across an image in the same manner as a convolution operation in a convolutional neural network (CNN). A digital projector and external lens form the desired 3×3 segment of an image on the ORN focal plane array. An oscilloscope measures the output V-t signal from a single node of the array, marked by V_{out} in Figure 3a. The output spectrum is then processed in software to obtain the FFT and filtered outputs. Figure 3b shows the I-V curves of all the SGM photodetectors in the experimental array under the same optical intensity (3 mW/mm^2). Figure 3c shows a representative V-t curve obtained from the 3×3 array when all the pixels are illuminated with uniform intensity. Figure 3d shows the frequency spectrum of the V-t curve of Figure 3c.

We then took the digital grayscale image of a cat (Fig. 2e) and projected it on the 3×3 ORN focal plane array, using the array as a convolution kernel with a stride of one (pixel intensity of 1 refers to 5.5 mW incident optical power). Figures 3e-h show the images obtained at 4 KHz ($\text{BW} = 100 \text{ Hz}$), 2.8 KHz ($\text{BW} = 200 \text{ Hz}$), 2.4 KHz ($\text{BW} = 800 \text{ Hz}$), and 3.2 KHz ($\text{BW} = 1.6 \text{ KHz}$), respectively. These filtered images demonstrate edge detection, intensity filtering, image sharpening and object segmentation operations. The circuit topology of this ORN kernel performs a multi-thresholding operation where the nonlinearly averaged intensity (\bar{P}) of the 3×3 pixels cell is mapped to a high value if $P_{low} < \bar{P} < P_{high}$ and to a low value if $\bar{P} < P_{low}$ or $\bar{P} > P_{high}$ where P_{low} and P_{high} changes with center frequency and bandwidth. As the bandwidth increases, $|P_{high} - P_{low}|$ becomes larger and can cover a larger range of pixel intensities. Therefore, at different frequencies, the ORN kernel thresholds the image within different pixel intensity ranges and the images shown in Figure 3e-h result from these different non-linear operations. As we increase the bandwidth from Figure 3e to 3h, we observe a larger image region thresholded to bright pixels. In this way, smaller bandwidth filters enable lower-level feature extraction, such as

edges, while high bandwidth filters lead to higher level feature extraction, such as object segmentation. Supplementary Figure S9 shows the processed images when the image is projected at different optical power ranges. When the incident optical power range is lower, similar image processing can be obtained at higher center frequencies. This result shows that the choice of optical power range is not very critical if appropriate center frequencies are chosen.

Next, we have investigated whether the same 3×3 ORN focal plane array can perform image segmentation from an image with multiple objects. A color image of size 180×156 pixels (Figure 3i) that features a Christmas tree, two dogs and a cat was selected. The image is split into three different grayscale images according to the pixel intensities of the color channels (R-channel, G-channel, and B-channel). Only pixels of the same color were coupled together. Therefore, each bandpass filter used had three different output images, one for each color channel. Figure 3j-m shows the images filtered at 3.0 KHz (B-channel), 3.5 KHz (B-channel), 4.0 KHz (B-channel), and 4.5 KHz (G-channel), respectively. The bandwidth used for each center frequency is 1 KHz. At 3.0 KHz (B-channel), the bright background emerges as white and rest of the image is thresholded to black, effectively segmenting the background. The images filtered at 3.5 KHz (B-channel) and 4 KHz (B-channel) segment the dog on the left and the cat, respectively. On the other hand, when filtered at 4.5 KHz (G-channel), the tree and the dog in the middle are detected. It is important to note that we have only used a single bandpass filter to segment an entire object in this case. Improved segmentation quality is expected when a linear combination of multiple frequencies is used. These results clearly illustrate how the ORN kernel can perform parallel, frequency multiplexed image processing and segmentation tasks.

These results show us two essential properties of this architecture: (1) the absence of any encoding or preprocessing for input, and (2) the ability to perform parallel computation at different

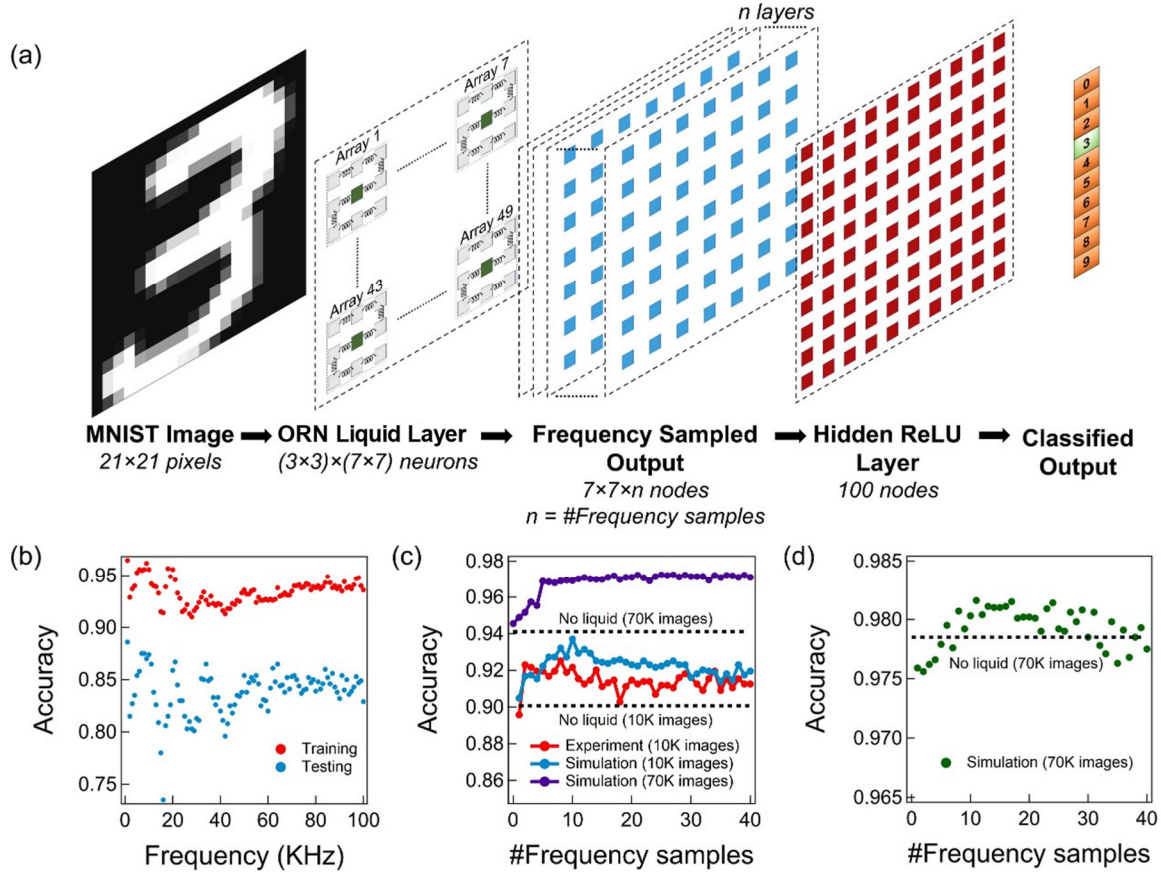


Figure 4: LSM implementation of ORN network for MNIST classification. (a) Image classification pipeline of the LSM structure showing an original input image, structure of the liquid layer, frequency sampled output images and further processing at the readout layer by hidden ReLU units. (b) Training and testing accuracy of the readout layer for training datasets corresponding to different frequency samples. (c) Classification accuracy of the handwritten digits as a function of number of frequency samples for 7×7 pixels/image and (d) for 21×21 pixels/sample.

frequencies. Since the projection of image and data acquisition are both performed in analog domain, inevitably noise is added to both the input and output of the system but can still obtain excellent results. It is also important to note that the circuit configuration used here to couple the oscillators is not unique. Supplementary Figure S10 shows image processing results obtained using a 10×10 coupled oscillator array kernel where the ORNs are connected to their nearest neighbors with a coupling inductance of 5 H. Engineering the circuit configurations allows the implementation of a variety of image processing functions.

Inference is carried out by using a 3×3 pixel coupled oscillator network to act as a liquid layer to construct a liquid state machine (LSM). Images from the MNIST database scaled to 21×21 pixels were serially projected on the 3×3 array with a stride of 3, while output signals were acquired from a single pixel. This data acquisition mode converts 21×21 images into $7\times 7\times n$ datapoints where n is the number of frequency samples considered. Each frequency sample corresponds to a bandpass filtered output at a given center frequency and a bandwidth of 1 KHz. Ten thousand images from the MNIST database were projected on the array, and the output data was collected and fed into a readout layer consisting of a single hidden layer with 100 nodes followed by a 10-node output layer. The hidden layer used a ReLU activation function, and the output layer used a softmax activation function. We use backpropagation to train only the readout layer while keeping the liquid layer connections untouched. Supplementary section S5 summarizes the implementation of the readout layer. Figure 4a shows the LSM schematic. Figure 4b plots the accuracy obtained at the 50th epoch if only a single frequency from each coupled array is fed into the hidden layer.

As expected, the single-frequency results show that the resulting accuracy varies by filter frequency. Feeding multiple frequency samples per pixel to the hidden layer is expected to augment the accuracy of the network. Figure 4c shows how feeding multiple frequencies into the hidden network modifies the testing accuracies obtained at the 200th epoch. We have done this for both experimental and simulated ORN arrays. The experiments were carried out on 10,000 images, limited by the speed of our data acquisition and projection setup. We observed a peak accuracy of 92.51% with 7 frequencies sampled per pixel. To evaluate the potential of this result if the full dataset of 70,000 images were used, a simulated version of the same 3×3 ORN focal plane array was also carried out. We see the resulting accuracy for the experimental and simulation cases with 10,000 images are very similar. As the simulation uses the experimental device I-V curves,

discrepancies between the simulation and experiment are attributed to the additional noise introduced by our image projection and data collection setup. The experimental data acquisition and simulation details are discussed in supplementary section S6. The in-sensor processed image performs better than the equivalent 7×7 input directly fed to the hidden layer without the liquid layer, which results in 90.06% accuracy. Similarly, through simulation we see that for all 70,000 images the accuracy reaches 97.21% with multiple frequencies, which is higher than the corresponding direct 7×7 data input (94.11%) into the neural network. Critically, if the 3×3 array is used as a convolution kernel with a stride of 1, a peak accuracy of 98.16% is achieved for 11 frequency samples per pixel, as shown in Figure 4d. This is higher than a standard imager directly inputting the 21×21 data (97.85%) into the neural network, showing improvement using this hardware over a purely software-defined approach. These results show that the parallel processing performed at different frequencies improves the network and that the coupling between pixels in the 3×3 array allows down-sampling of the number of outputs to the hidden layers of the fully connected neural network. In addition, the LSM architecture does not require the training of liquid layer interconnections, which significantly reduces the complexity and computational cost of the training.

While an ORN array does not require any external electrical power to drive the oscillations, the system requires peripheral circuitry to read the voltages and perform bandpass filtering operations. A charge domain on-chip FFT processor⁵⁵ can perform such operations with a low energy cost. As discussed in supplementary section S7, an ORN array can perform convolution equivalent tasks with a performance of 42211 TOPS/W, which translates to an energy cost of 24 aJ/OP with a precision equivalent to 8-bit integer operations in digital systems. These projections clearly show that frequency multiplexed computing using coupled ORN array has the potential to

228 completely replace the energy-expensive convolutional layers in CNN for deep learning
 229 applications.

NPU application	Type	Comment	bits	Reported TOPS/W	Normalized TOPS/W (8 bits)
Analog to information conversion ¹	Analog	In-sensor NN	8	43.5	43.5
VMM ²	Digital	SRAM	4	351	87.75
MAC macro ¹⁶	Analog	DRAM	4	217	54.25
VMM ²³	Digital	DNN learning processor	8	146.52	146.52
Arithmetic logic ³¹	Digital	Superconducting logic devices	8	120	120
VMM ³²	Analog	Si-CMOS/CAAC-IGZO based memory	6	210	118.13
VMM ⁴⁰	Digital	Stochastic NN accelerator	8	75	75
MAC macro ⁵²	Digital	SRAM	8	63	63
MAC macro ⁵³	Analog	SONOS memory	8	100	100
MAC macro ⁵⁴	Digital	SRAM	1	20943	327.23
General purpose	Digital	NVIDIA A100	8	4.992	4.992
General purpose	Digital	Apple a16 Bionic	8	2.67	2.67
General purpose	Digital	Qualcomm Snapdragon 865	8	4.5	4.5
Nonlinear convolution (3×3 kernel) (This work)	Analog	ORN	8	42211	42211

Table 1: Comparison between different NPUs

230 Table 1 shows the performance comparison between different neural processing units
 231 (NPU) for deep learning. Different NPUs operate at different bit resolutions and therefore an n-bit

performance was scaled by a factor of $\left(\frac{n}{8}\right)^2$ to get a normalized 8-bit performance. Such a scaling is reasonable^{56,57} since number of transistors in digital logic typically scales as $\sim n^2$.

In conclusion, we have introduced in-sensor neuronal computing as an alternative to in-sensor synaptic computing. We demonstrated that coupled ORNs enable highly parallel, frequency multiplexed computation on input images without data conversion, storage, or transmission penalties. Experimental implementations using 3×3 arrays of coupled ORNs show parallel image processing on projected images. These include edge detection, intensity filtering, and object segmentation as examples of image processing tasks carried out at the detector array. We have also demonstrated that inference with these devices performs handwritten digit classification from the MNIST database with higher accuracy than traditional photodetectors. While we have focused on image classification and image processing applications, we expect this computational approach to be general. Most importantly, ORN-based computation is extremely energy efficient with an estimated performance of ~ 42211 TOPS/W considering the energy cost of the peripheral circuits, laying the framework for a general, ultralow power, variation tolerant approach to oscillator-computing.

Methods

Semiconductor substrate preparation. Moderately boron doped ($N_A = 5\times 10^{15} \text{ cm}^{-3}$) silicon (100) wafer was used as the semiconductor substrate. A 5 nm Ti/60 nm Au mesh is photolithographically defined and deposited by electron beam evaporation. A monolayer of CVD grown graphene is transferred on top of the metal mesh via wet transfer method⁵⁸. A 100 nm aluminum film sputtered at the back side of the substrate acts as the contact to silicon.

Graphene growth and transfer. CVD graphene was grown on a Cu foil by using low pressure CVD. Cu foil was etched inside FeCl_3 copper etchant for 30 seconds before the graphene growth. Cu foil was annealed in a quartz tube furnace at 1000°C for 30 min with 50 standard cubic centimeters per minute (sccm) hydrogen (H_2) flow rate. Graphene was synthesized under 7 sccm of methane (CH_4) and 50 sccm of hydrogen (H_2) for 40 min. For transfer, Poly(methyl methacrylate) (PMMA A6495) was spin-coated on top of Cu foil at 2000 rpm for 60 sec and baked for 5 min under 170°C . PMMA spin-coated Cu foil was etched using FeCl_3 copper etchant graphene to remove the Cu while the remaining PMMA/Graphene floats to the top. The stacked layer was cleaned with D.I water and transferred to 10% hydrochloric acid solution to remove the remaining Cu etchants. After cleaning with D.I water once more, PMMA/Graphene was transferred on top of the oxide/semiconductor substrate. The substrate was dried in the air overnight followed by 90°C for 15min, 150°C for 30min, and 90°C for 15min to ensure the adhesion between the graphene and the substrate. Finally, the substrate was immersed in acetone for 12 hours to remove the PMMA.

Raman spectroscopy for graphene. CVD grown monolayer graphene transferred on the substrate was analyzed by Raman spectroscopy. Raman spectra were collected with Renishaw spectrometer with a 532-nm laser focused in a $0.5\text{-}\mu\text{m}$ spot through a Leica microscope with a 100x objective lens.

Wavelength dependent measurements. A supercontinuum laser with grating monochromator was used to illuminate the SGM photodetector with lights of different wavelengths between 400 and 1100 nm. Applied voltage was stepped while light and dark current measurements were performed. The difference between these two current measurements, i.e., the photocurrent was then used to measure the responsivity of the device.

277 **ORN measurements.** A 5×5 array of SGM photodetectors was fabricated and individual devices
278 were wirebonded to a PCB. The devices were electrically connected to the inductors (all 10 mH)
279 on a breadboard to form the ORN kernel. A digital projector was used to project the patterns on
280 the device array (a 3×3 array from the 5×5 array) and an oscilloscope was used to record the
281 oscillation waveforms. The whole process was automated using MATLAB environment.

282 **Acknowledgments:**

283 This work was supported by Department of Energy Grant No. DE-SC0022248, National Science
284 Foundation Award No. 2004791, Office of Naval Research Grant No. N00014-21-1-2634, Air
285 Force Office of Scientific Research Grant No. FA9550-21-1-0305. R.A. and J.T. acknowledge
286 USC Provost Graduate Fellowships. S.D. acknowledges USC Graduate School Fellowship.

287 **Author contributions:**

288 R.A. and R.K. conceived the project and designed the experiments. R.A, H.U.C, S.D, and J.T.
289 performed the device fabrication. H.U.C, J.T., and S.D. carried out graphene growth. H.L. and
290 J.W. performed the temperature-dependent measurements. R.A. performed the measurements and
291 simulations. R.A. and S.A.A.J designed and implemented the ORN measurement setup. All
292 authors contributed to analyzing the data. R.A. and R.K. wrote the paper while all the authors
293 provided feedback.

294 **Data availability:**

295 The data that support the plots within this paper and other findings of this study are available from
296 the corresponding author upon reasonable request.

297

298 **Code availability:**

299 The codes used to perform the simulations are available from the corresponding author upon
300 reasonable request.

301 **Competing interests:**

302 The authors declare no competing interests.

303

304 **References:**

- 305 1 Sadasivuni, S., Bhanushali, S. P., Banerjee, I. & Sanyal, A. In-sensor neural network for high energy
306 efficiency analog-to-information conversion. *Scientific reports* **12**, 18253 (2022).
- 307 2 Dong, Q. *et al.* in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*. 242-244 (IEEE).
- 308 3 Amir, M. *et al.* in *2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*. 1-2 (IEEE).
- 309 4 Amir, M. F., Ko, J. H., Na, T., Kim, D. & Mukhopadhyay, S. 3-D stacked image sensor with deep
310 neural network computation. *IEEE Sensors Journal* **18**, 4187-4199 (2018).
- 311 5 Choo, K. D. *et al.* Energy-efficient motion-triggered IoT CMOS image sensor with capacitor array-
312 assisted charge-injection SAR ADC. *IEEE Journal of Solid-State Circuits* **54**, 2921-2931 (2019).
- 313 6 Du, Z. *et al.* in *Proceedings of the 42nd Annual International Symposium on Computer Architecture*.
314 92-104.
- 315 7 Finateu, T. *et al.* in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*. 112-114 (IEEE).
- 316 8 Hsu, T.-H. *et al.* in *2019 IEEE International Electron Devices Meeting (IEDM)*. 22.25. 21-22.25. 24
317 (IEEE).
- 318 9 Hsu, T.-H. *et al.* A 0.8 V multimode vision sensor for motion and saliency detection with ping-pong
319 PWM pixel. *IEEE Journal of Solid-State Circuits* **56**, 2516-2524 (2021).
- 320 10 Jiménez-Fernández, A. *et al.* A binaural neuromorphic auditory sensor for FPGA: A spike signal
321 processing approach. *IEEE transactions on neural networks and learning systems* **28**, 804-818
322 (2016).
- 323 11 Lichtsteiner, P., Posch, C. & Delbruck, T. A 128×128 120 dB $15 \mu s$ latency
324 asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits* **43**, 566-576
325 (2008).
- 326 12 LiKamWa, R., Hou, Y., Gao, J., Polansky, M. & Zhong, L. Redeye: analog convnet image sensor
327 architecture for continuous mobile vision. *ACM SIGARCH Computer Architecture News* **44**, 255-
328 266 (2016).
- 329 13 Wang, C.-Y. *et al.* Gate-tunable van der Waals heterostructure for reconfigurable neural network
330 vision sensor. *Science Advances* **6**, eaba6173 (2020).
- 331 14 Zhou, F. & Chai, Y. Near-sensor and in-sensor computing. *Nature Electronics* **3**, 664-671 (2020).
- 332 15 Mennel, L. *et al.* Ultrafast machine vision with 2D material neural network image sensors. *Nature*
333 **579**, 62-66 (2020).
- 334

335 16 Chen, Z., Chen, X. & Gu, J. in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*. 240-
336 242 (IEEE).

337 17 Lyon, R. F. & Mead, C. An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech, and*
338 *Signal Processing* **36**, 1119-1134 (1988).

339 18 Hasler, P., Smith, P. D., Graham, D., Ellis, R. & Anderson, D. V. Analog floating-gate, on-chip
340 auditory sensing system interfaces. *IEEE Sensors Journal* **5**, 1027-1034 (2005).

341 19 Hsieh, H.-Y. & Tang, K.-T. VLSI implementation of a bio-inspired olfactory spiking neural network.
342 *IEEE transactions on neural networks and learning systems* **23**, 1065-1073 (2012).

343 20 Wen, B. & Boahen, K. in *2006 IEEE International Solid State Circuits Conference-Digest of Technical*
344 *Papers*. 2268-2277 (IEEE).

345 21 Ellis, R., Yoo, H., Graham, D. W., Hasler, P. & Anderson, D. V. in *2002 IEEE International Symposium*
346 *on Circuits and Systems. Proceedings (Cat. No. 02CH37353)*. II-II (IEEE).

347 22 Schrauwen, B., D'Haene, M., Verstraeten, D. & Van Campenhout, J. 1097-1102 (IEEE).

348 23 Kim, S., Lee, J., Kang, S., Lee, J. & Yoo, H.-J. in *2020 IEEE Symposium on VLSI Circuits*. 1-2 (IEEE).

349 24 Tan, H. *et al.* Tactile sensory coding and learning with bio-inspired optoelectronic spiking afferent
350 nerves. *Nature communications* **11**, 1-9 (2020).

351 25 Kim, Y. *et al.* A bioinspired flexible organic artificial afferent nerve. *Science* **360**, 998-1003 (2018).

352 26 Zhang, X. *et al.* An artificial spiking afferent nerve based on Mott memristors for neurorobotics.
353 *Nature communications* **11**, 1-9 (2020).

354 27 Zhou, F. *et al.* Optoelectronic resistive random access memory for neuromorphic vision sensors.
355 *Nature nanotechnology* **14**, 776-782 (2019).

356 28 Zhang, X. & Basu, A. in *2022 IEEE Custom Integrated Circuits Conference (CICC)*. 1-2 (IEEE).

357 29 Bose, S. K. & Basu, A. in *2021 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. 1-3 (IEEE).

358 30 Chai, Y. (Nature Publishing Group UK London, 2020).

359 31 Nagaoka, I. *et al.* in *2019 IEEE International Superconductive Electronics Conference (ISEC)*. 1-3
360 (IEEE).

361 32 Chen, M.-C. *et al.* in *2022 International Electron Devices Meeting (IEDM)*. 18.12. 11-18.12. 14
362 (IEEE).

363 33 Chou, J., Bramhavar, S., Ghosh, S. & Herzog, W. Analog coupled oscillator based weighted Ising
364 machine. *Scientific reports* **9**, 14786 (2019).

365 34 Dutta, S. *et al.* An Ising Hamiltonian solver based on coupled stochastic phase-transition nano-
366 oscillators. *Nature Electronics* **4**, 502-512 (2021).

367 35 Wang, T. & Roychowdhury, J. in *Unconventional Computation and Natural Computation: 18th*
368 *International Conference, UCNC 2019, Tokyo, Japan, June 3–7, 2019, Proceedings* **18**. 232-256
369 (Springer).

370 36 Nikonov, D. E. *et al.* Convolution inference via synchronization of a coupled CMOS oscillator array.
371 *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* **6**, 170-176 (2020).

372 37 Delacour, C., Carapezzi, S., Abernot, M. & Todri-Sanial, A. Energy-Performance Assessment of
373 Oscillatory Neural Networks based on VO2 Devices for Future Edge AI Computing. (2022).

374 38 Núñez, J. *et al.* Oscillatory Neural Networks Using VO2 Based Phase Encoded Logic. *Frontiers in*
375 *Neuroscience* **15** (2021). <https://doi.org/10.3389/fnins.2021.655823>

376 39 Corti, E. *et al.* Coupled VO2 oscillators circuit as analog first layer filter in convolutional neural
377 networks. *Frontiers in Neuroscience* **15**, 628254 (2021).

378 40 Romaszkan, W. *et al.* A 4.4–75-TOPS/W 14-nm Programmable, Performance-and Precision-
379 Tunable All-Digital Stochastic Computing Neural Network Inference Accelerator. *IEEE Solid-State*
380 *Circuits Letters* **5**, 206-209 (2022).

381 41 Nagumo, J., Arimoto, S. & Yoshizawa, S. An active pulse transmission line simulating nerve axon.
382 *Proceedings of the IRE* **50**, 2061-2070 (1962).

- 42 Smith-Saville, R. Variable frequency tunnel diode relaxation oscillator. *Nuclear Instruments and Methods* **55**, 120-124 (1967).
- 43 Keener, J. P. Analog circuitry for the van der Pol and FitzHugh-Nagumo equations. *IEEE transactions on systems, man, and cybernetics*, 1010-1014 (1983).
- 44 Izhikevich, E. M. & FitzHugh, R. Fitzhugh-nagumo model. *Scholarpedia* **1**, 1349 (2006).
- 45 Ho, C.-L., Wu, M.-C., Ho, W.-J. & Liaw, J.-W. Light-induced negative differential resistance in planar InP/InGaAs/InP double-heterojunction p-i-n photodiode. *Applied Physics Letters* **74**, 4008-4010 (1999). <https://doi.org/10.1063/1.123243>
- 46 Liu, W. *et al.* Light-induced negative differential resistance in gate-controlled graphene-silicon photodiode. *Applied Physics Letters* **112**, 201109 (2018). <https://doi.org/10.1063/1.5026382>
- 47 Qin, S. *et al.* A light-stimulated synaptic device based on graphene hybrid phototransistor. *2D Materials* **4**, 035022 (2017).
- 48 Wang, X. *et al.* Light-induced negative differential resistance effect in a resistive switching memory device. *Current Applied Physics* **20**, 371-378 (2020). <https://doi.org/10.1016/j.cap.2019.12.008>
- 49 Antonova, I. V., Shojaei, S., Sattari-Esfahlan, S. & Kurkina, I. I. Negative differential resistance in partially fluorinated graphene films. *Applied Physics Letters* **111**, 043108 (2017).
- 50 Zhang, Q. *et al.* Negative differential resistance and hysteresis in graphene-based organic light-emitting devices. *Journal of Materials Chemistry C* **6**, 1926-1932 (2018).
- 51 Lee, K. W. *et al.* Light-induced negative differential resistance in graphene/Si-quantum-dot tunneling diodes. *Scientific Reports* **6**, 30669 (2016).
- 52 Fujiwara, H. *et al.* in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*. 1-3 (IEEE).
- 53 Agrawal, V. *et al.* in *2020 IEEE International Memory Workshop (IMW)*. 1-4 (IEEE).
- 54 Lin, C.-S. *et al.* in *2021 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. 1-3 (IEEE).
- 55 Sadhu, B., Sturm, M., Sadler, B. M. & Harjani, R. Analysis and design of a 5 GS/s analog charge-domain FFT for an SDR front-end in 65 nm CMOS. *IEEE Journal of Solid-State Circuits* **48**, 1199-1211 (2013).
- 56 Yang, Q. & Li, H. BitSystolic: A 26.7 TOPS/W 2b~ 8b NPU with configurable data flows for edge devices. *IEEE Transactions on Circuits and Systems I: Regular Papers* **68**, 1134-1145 (2020).
- 57 Horowitz, M. in *IEEE, feb.*
- 58 Rezaeifar, F., Ahsan, R., Lin, Q., Chae, H. U. & Kapadia, R. Hot-electron emission processes in waveguide-integrated graphene. *Nature Photonics*, 1-6 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryORNnaturesubmission.pdf](#)