Supplementary Note

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

Ragib Ahsan[1], Hyun Uk Chae[1], Seyedeh Atiyeh Abbasi Jalal[1], Zezhi Wu[1], Jun Tao[1], Subrata Das[1], Hefei Liu[1], Jiang-Bin Wu[1], Stephen Cronin[1], Han Wang[1], Constantine Sideris[1], Rehan Kapadia[1, *]

[1]Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA

*Email: rkapadia@usc.edu

S1. NDR mechanism

S1.1 Control devices

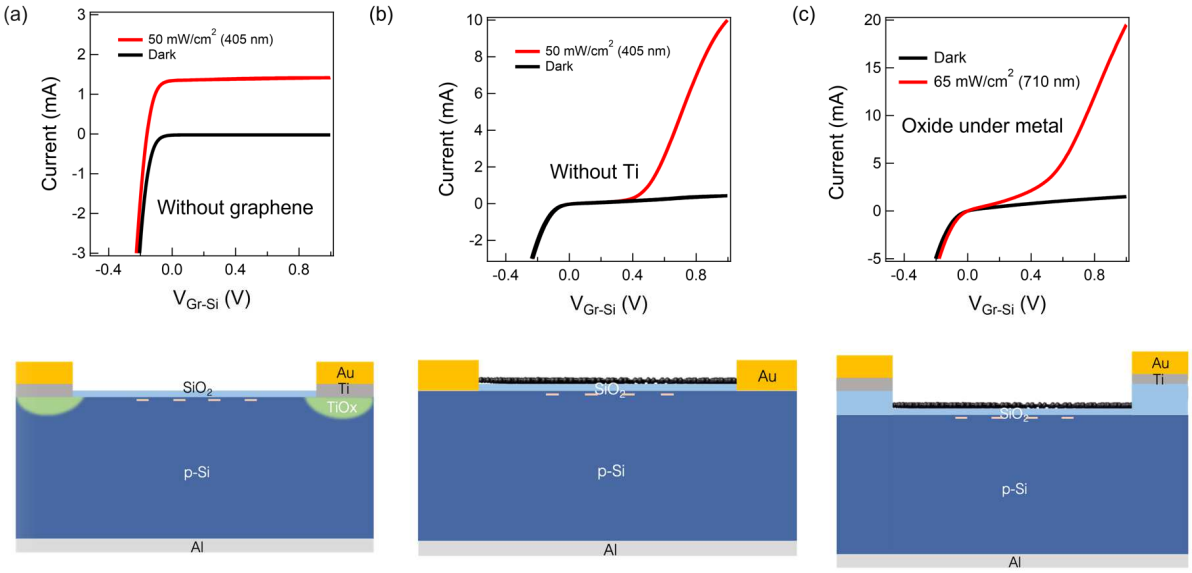


Figure S1: I-V characteristics of control devices. (a) I-V curves in dark and light conditions for a device without graphene, (b) a device with graphene but without Ti in the metal, and (c) a device with graphene and Ti/Au metal mesh but thick oxide underneath the metal mesh.

To elucidate the mechanism of NDR, we have performed I-V measurements on three different control devices: (1) a silicon-Ti/Au grid device without graphene, (2) a silicon-Au grid-graphene device, and (3) a silicon-Ti/Au grid-graphene device with 200 nm thick PECVD SiO_2 under the Ti/Au mesh only. Supplementary Figure S1 shows the I-V curves of these three devices, where we observe that the NDR behavior is absent for all three. The device without graphene operates only in the photovoltaic regime, while the device without Ti and the device with thick oxide under metal do not operate in the photovoltaic regime. These behaviors indicate that photovoltaic regime of operation is directly coming from the collection of charge carriers by the metal mesh itself and the Ti/Si Schottky barrier is crucial to the device operation. The graphene generates the NDR, which is responsible for the decrease in the collected photocurrent as a function of applied bias.

S1.2 NDR Model

Figure S2 shows the schematic for our proposed model for the observed NDR behavior in this device. Two possible channels can collect the photogenerated electrons: (1) lateral diffusion in the plane of the silicon surface to reach the Ti/Au contact or (2) collection into graphene through the thin native oxide barrier. These two collection channels compete with recombination processes at

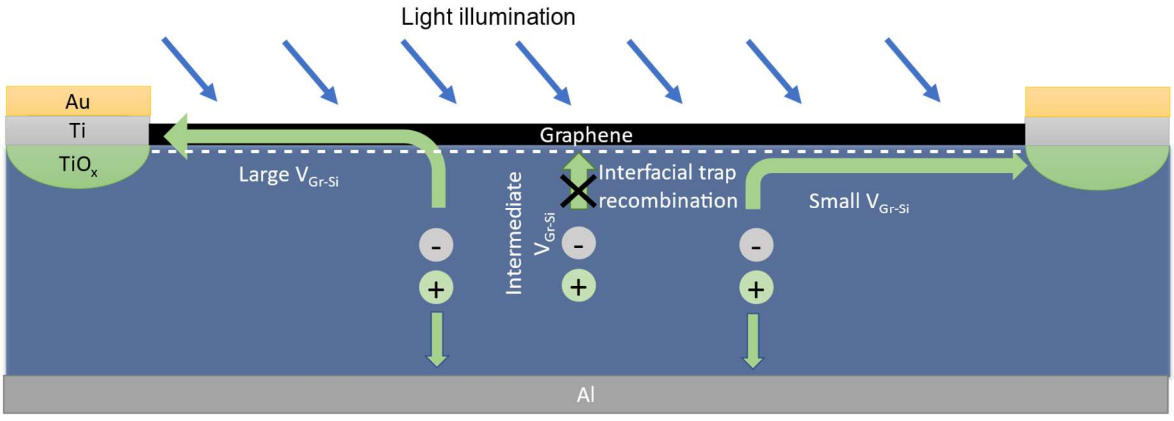


Figure S2: Schematic of the NDR mechanism showing the competing transport channels for the photogenerated carriers.

the Si/graphene interface and bulk. The Ti/Au contact region collects the majority of electrons at lower voltages when the native oxide barrier is opaque. When the voltage is increased, the barrier between the graphene and native oxide barrier becomes less opaque, collecting more photogenerated electrons in the graphene. However, the surface Fermi level will also move, modifying the density of unoccupied interface defect states, which modifies the interfacial recombination rates. Further increasing the voltage causes the native oxide barrier to become transparent and the defect states to fill. The electrons tunnel into graphene and overall recombination reduces(photoconductive regime).

S1.3. Position and temperature dependent measurements

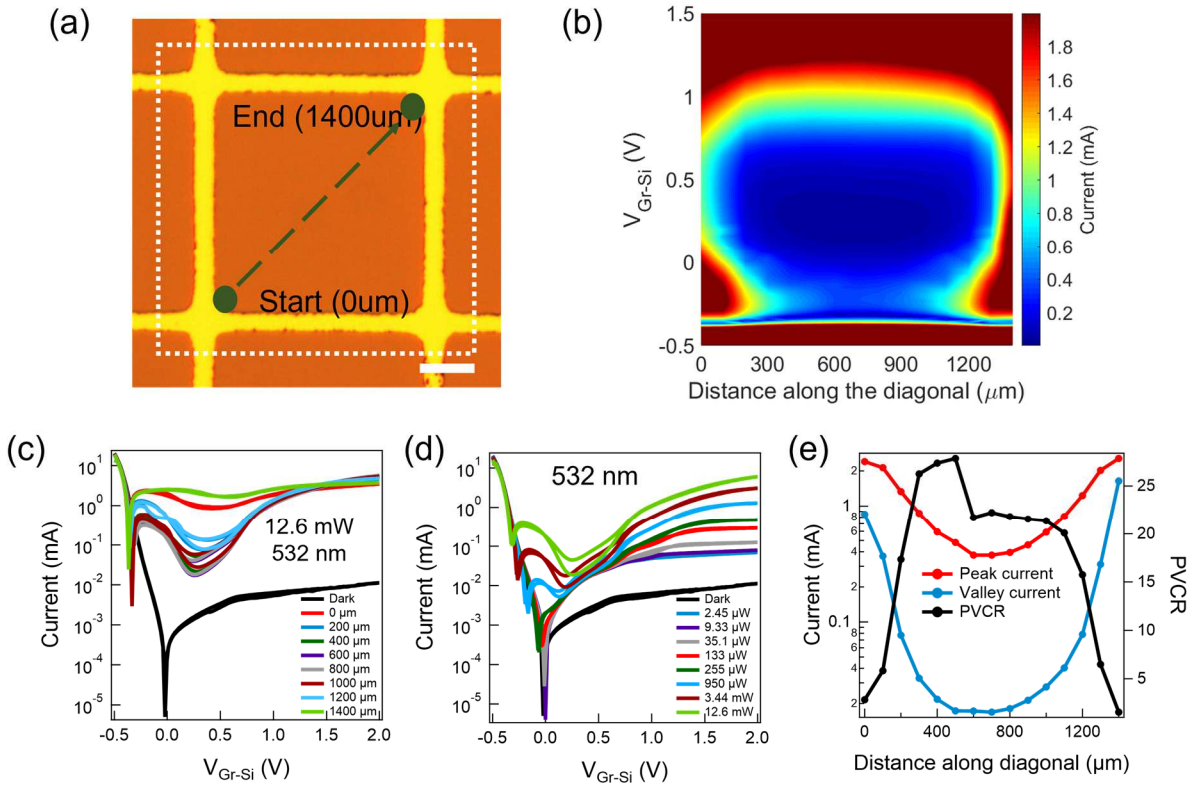


Figure S3: (a) Optical micrograph of the grid in the device showing direction of position dependent measurement. (b) Spatial dependence of current for focused beam measurements at 532 nm wavelength at a power of 12.6 mW. (c) I-V curves corresponding to the colormap produced in (b). (d) I-V curves measured at different optical powers for beam position at the midpoint of the diagonal. (e) Spatial dependence of peak and valley current and PVCR.

To verify the proposed mechanism, we have performed a position dependent I-V measurement on the device using a focused laser beam of 532 nm wavelength with a spot size of ~100 µm^2 and power of 12.6 mW. We have scanned the beam through the diagonal of a single mesh square, as shown in Supplementary Figure S3a. The device has a square mesh with 20 um width and 1 mm pitch. The beam was first positioned at one of the intersections of the mesh and then moved through the diagonal. Since each square has 1 mm sides, the diagonal length is 1.414 mm. As the beam

moves through the diagonal, it reaches the maximum distance from the metal mesh when it reaches the midpoint of the square diagonal, i.e., 0.707 mm from the intersection. We observe a change in I-V characteristics as a function of the distance between the beam and the metal grid. Due to the square geometry of the mesh, the distance from the contacts is maximum when the beam is at the midpoint of the diagonal. Supplementary Figure S3b shows a colormap of the current magnitude in as we sweep the voltage and the distance along the mesh diagonal. For $V_{Gr\text{-}Si} < 1$ V, the current varies with distance along the diagonal. The I-V curves used to produce this colormap is shown in supplementary Figure S3c. Supplementary Figure S3d shows the I-V curves measured at the midpoint of the mesh diagonal for different optical power densities where we can observe a threshold optical power required to obtain the NDR behavior. Supplementary Figure S3e shows a line plot of the peak, valley, and peak-to-valley current ratio (PVCR) as a function of the distance along the mesh diagonal. The peak PVCR occurs when light is incident in the middle of the diagonal because of the large relative change in valley current.

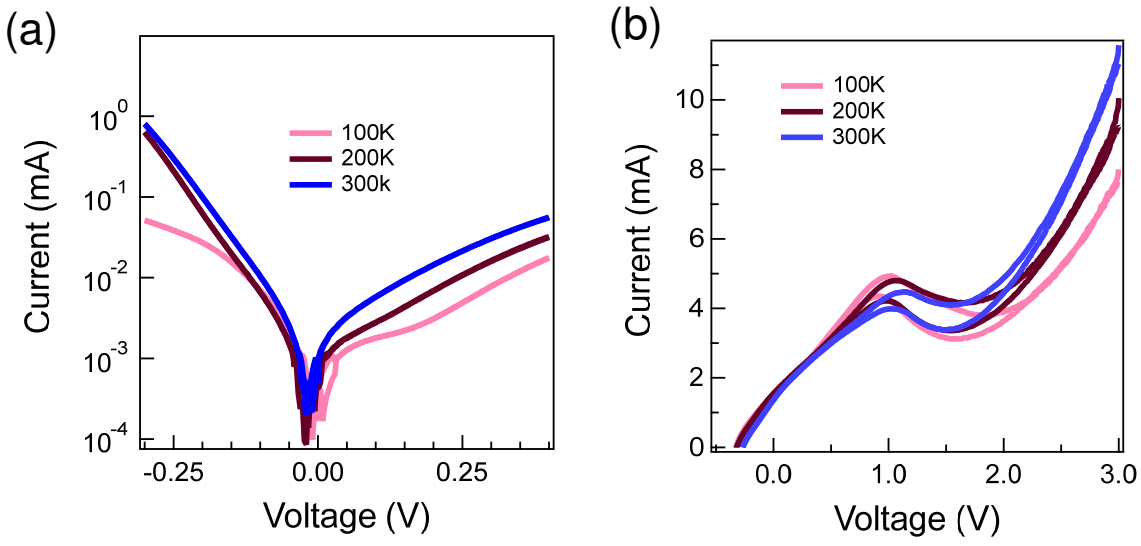


Figure S4: Temperature dependent I-V curves in (a) log and (b) linear scale.

Temperature dependent measurements were performed on the device with a diffused laser beam of 635 nm wavelength at 40 mW optical power. Temperature of the device was varied from 80K to 300K. The temperature dependent I-V curves (supplementary Figure S4) show that the NDR behavior does not have a significant dependence on temperature, consistent with the proposed model.

S1.4. Capacitance measurements

Capacitance of the device has been measured under both dark and illumination conditions using a capacitance measurement unit of Keysight B1500a semiconductor parameter analyzer. The measurements were performed under different optical power densities at 445 nm wavelength and for different measurement frequencies (1 KHz, 10 KHz, 100 KHz and 1 MHz). Supplementary Figure S5a shows the C-V curves for the device measured at a small signal frequency of 1 kHz

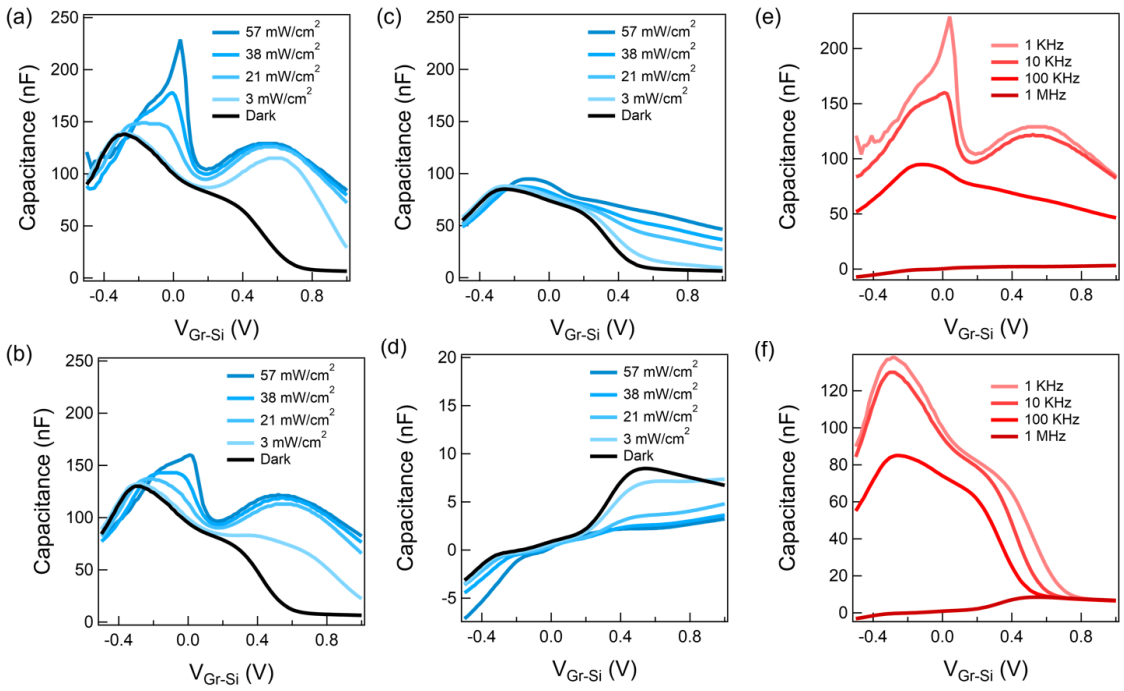


Figure S5: Capacitance measurements of NDR device. Small signal capacitance-voltage characteristics of NDR device at different power densities for (a) 1 KHz, (b) 10 KHz, (c) 100 KHz, and (d) 1 MHz. (e) Small signal capacitance measured at a power density of 57 mW/cm^2 and (f) dark conditions for different frequencies

under uniform optical illumination with varied power densities at 445 nm wavelength. The dark C-V curve shows an initial increase in capacitance due to formation of a depletion region and then a decrease in the capacitance as the width of depletion region increases with increasing reverse bias voltage. The C-V behavior under illumination shows a larger initial capacitance followed by a sharper decrease in capacitance for lower voltages. The larger capacitance under illumination can be attributed to the increase in charge in the depletion region due to photogenerated carriers while the sharp decrease can be attributed to the presence of a recombination process that annihilates these photogenerated carriers. When the voltage is increased further, we see another slow increase followed by a slow decrease in capacitance unlike the dark measurements. When we compare these C-V measurements with the higher frequency C-V measurements (supplementary Figure S5b-f), we observe that the second increase in the capacitance disappears at the higher frequencies. This frequency dependent behavior supports the hypothesis that charges responsible for the second increase in capacitance are due to slow charge trapping centers in the device. These experimental results agree with the mechanism proposed in supplementary Figure S2.

S1.5. Sentaurus simulations

To further validate the proposed mechanism, we have performed TCAD Sentaurus simulations of the device. We have simulated the carrier transport characteristics of the device using a drift-diffusion model in Sentaurus TCAD. For the simulations, we have considered a 2D device with 100 um width instead of the 1 mm width of the real device to save simulation resources. Supplementary Figure S6a shows a schematic of the simulated device. The device has 100 um thick 5e15 cm^{-3} boron doped p-Si as substrate and a 3 nm thick layer of SiO_2 on top of silicon. We have used a 1e17 cm^{-3} phosphorous doped n-Si as proxy to the TiO_x layer underneath the metal mesh. Then we have used a 1 nm thick semiconductor with 0.01 eV bandgap and an electron

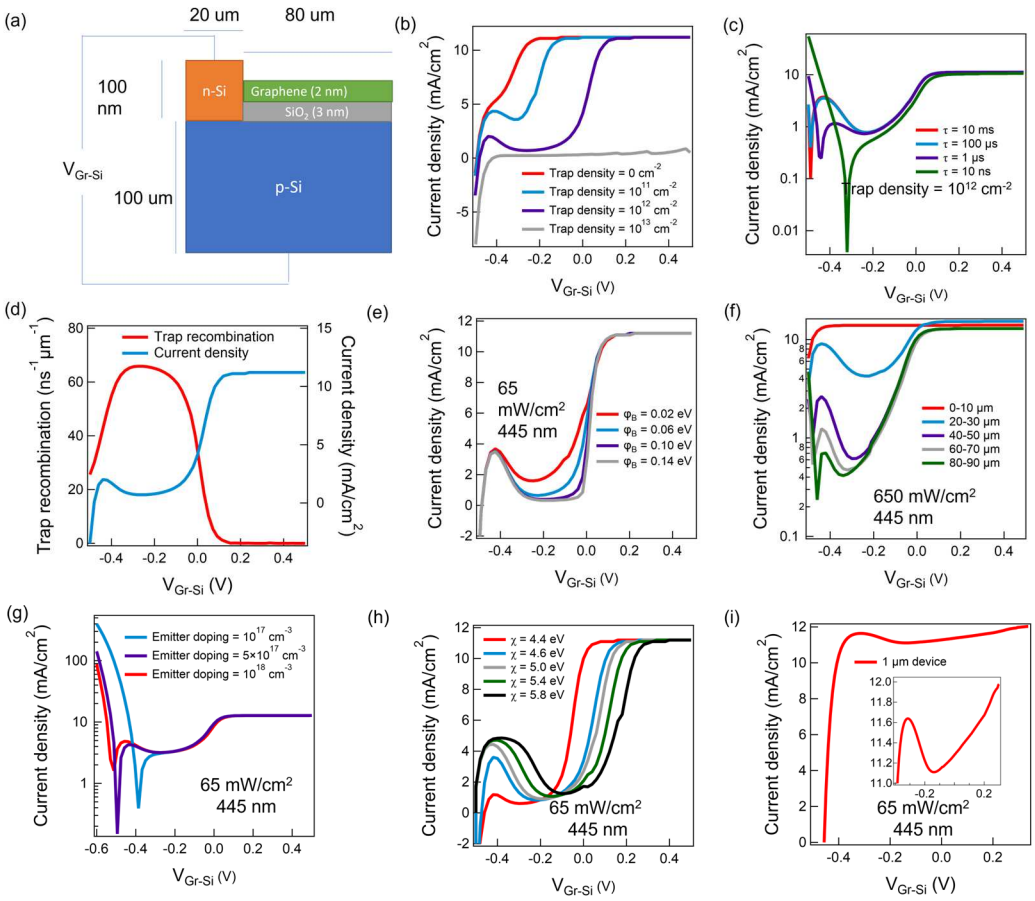


Figure S6: Sentaurus simulations of the NDR device. (a) Schematic of the simulated device (b) Simulations showing the effect of charge traps and (c) carrier lifetime in determining NDR behavior for a trap density of 10^{12} cm^{-2}. (d) Sentaurus simulation showing the increased recombination at the trap states in the NDR regime for trap density of 10^{12} cm^{-2}. (e) Modification of the valley current with barrier height (f) Position dependent illumination and corresponding I-V sweeps (g) Change in V_{OC} with n-Si doping (h) Modification of NDR behavior as electron affinity of the graphene layer was changed, (i) Scalability of the NDR device: I-V curve of a 1 μm device showing NDR behavior.

affinity of 4.5 eV as a proxy to graphene layer. We have considered a direct tunneling mechanism between silicon and graphene with adjustable tunnel barrier height. Supplementary Figure S6b shows the J-V curves calculated for different electron trap densities under illumination of a power density of 65 mW/cm^2 at 445 nm wavelength. Tunnel barrier was fixed at 50 meV for these calculations. As the trap density is increased, we can see the emergence of NDR behavior which again vanishes when the trap density becomes too large. Supplementary Figure S6c shows the effect of bulk electron lifetime in silicon on the NDR behavior for a trap density of 10^{12} cm^{-2}.

When the lifetime is short (10 ns), the bulk recombination dominates over the interfacial recombination, and therefore no NDR is observed. However, NDR is observed when carriers have a long lifetime (>100 μs) in the bulk and interfacial recombination becomes more prominent. Supplementary Figure S6d shows the recombination rate at the electron trap states for a trap density of 10^{12} cm^{-2} as well as the J-V curve. We can clearly see increase in recombination rate at the trap states when the NDR regime starts and then we see a subsequent decrease in recombination rate as NDR regime ends and current starts to increase.

Then we have fixed the electron trap density at 10^{12} cm^{-2} and varied the tunnel barrier height as shown in supplementary Figure S6e. We can clearly observe the increasing sharpness of the I-V curves as the tunnel barrier is increased. Then we have simulated for position dependent illumination with 10 um spot width (supplementary Figure S6f). When the light is shined on the n-Si region (0-10 um), we cannot observe any NDR behavior. This is consistent with what we observe experimentally where we also do not observe NDR when light is shined very close to the metal mesh. As we move further from the n-Si region, we see a decrease in the current which is also consistent with experimental observations. Since n-Si is used as a proxy to the TiO_x layer, we have varied the doping in n-Si to investigate how that affects the behavior of the device as shown in supplementary Figure S6g. When the doping is increased, there is an increased electric field at the junction which causes more electrons to be collected and increases the open circuit voltage of the device. We have also varied the electron affinity of the graphene layer to see how it affects the NDR behavior as shown in supplementary Figure S6h. An increasing electron affinity causes an increased electric field at the depletion region at the silicon/native oxide/graphene interface due to larger difference in electron affinity of silicon and graphene. As a result, the collection of electrons at the depletion region improves and current increases in photovoltaic regime. However, when the

depletion region has a strong electric field, there is a strong pull on the electrons toward the n-Si region which causes the NDR regime to occur at even higher voltages. In order to investigate the scalability of the device, we have performed simulation on a device with 1 um width as well. As shown in supplementary Figure S6i, the 1 um device also shows NDR behavior when illuminated with 65 mW/cm^2 optical power density at 445 nm wavelength. However, it will be impractical to scale the devices below 1 um because of diffraction. While our experimental results show that the charge trapping states at the silicon/oxide interface are responsible for the NDR behavior, the TCAD simulations help us quantitatively verify the validity of the proposed mechanism.

S1.6. Responsivity measurements

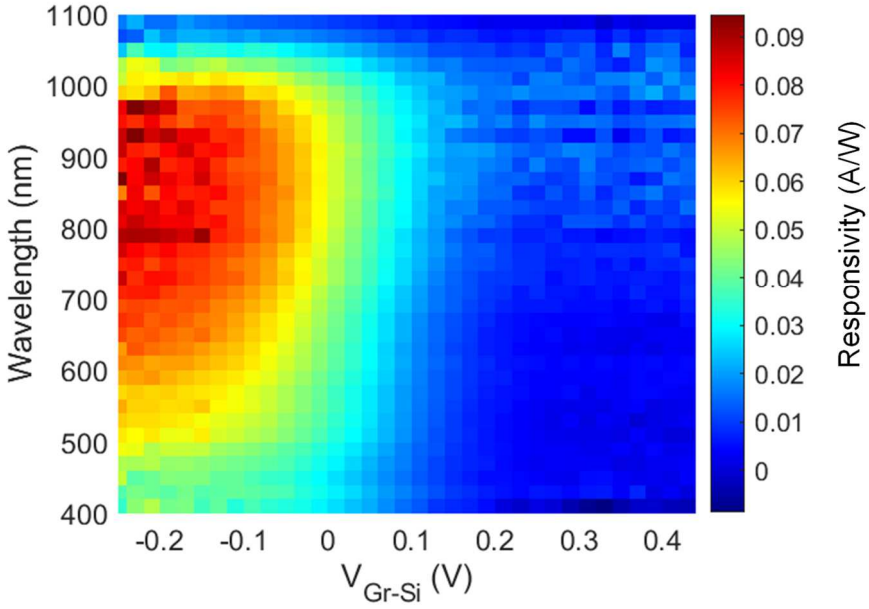


Figure S7: Wavelength dependent responsivity measurements

S2. Functions implemented by a single ORN

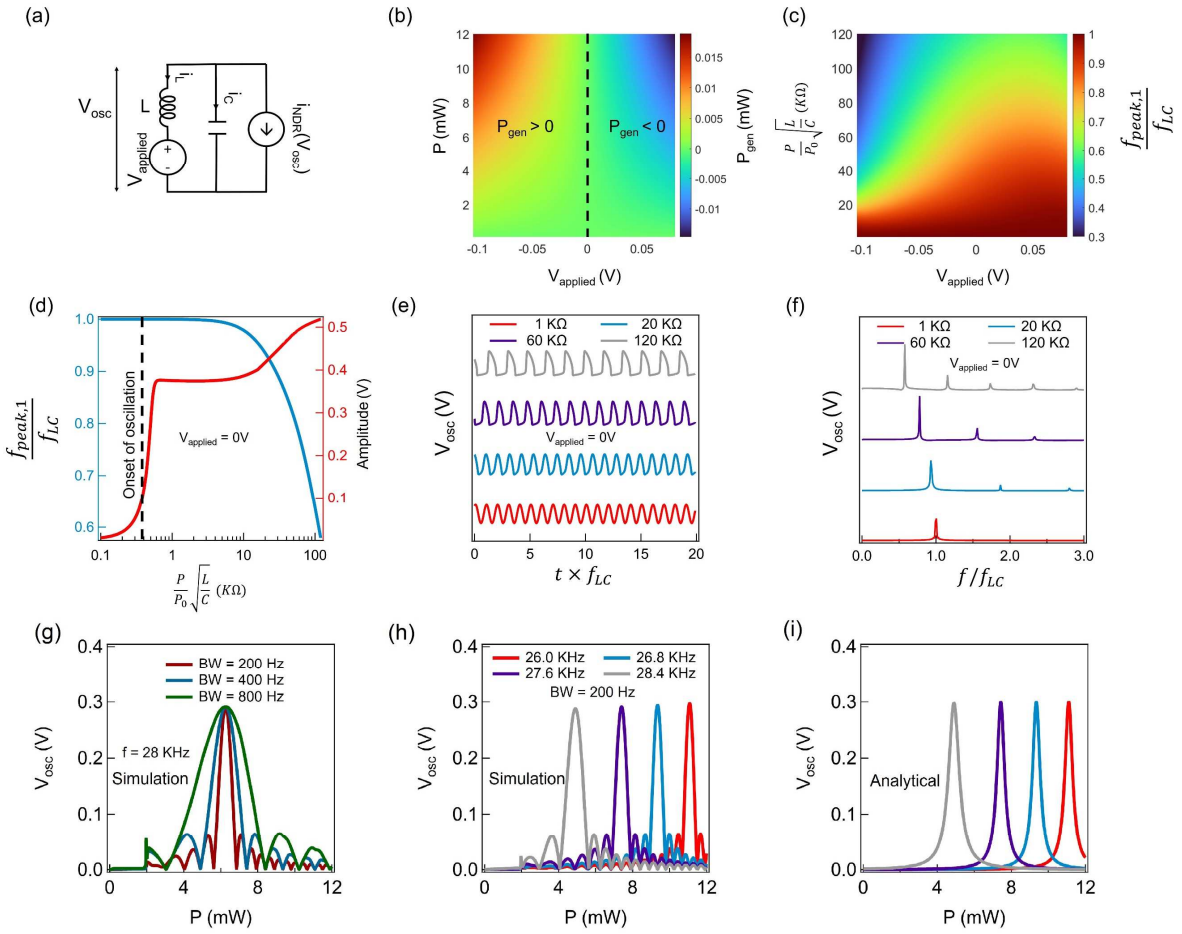


Figure S8: (a) Schematic of a single ORN (b) Power generated by ORN (c) Normalized fundamental frequency of oscillation as a function of normalized optical power and applied voltage (d) Normalized fundamental frequency and amplitude at $V_{applied} = 0V$ (e) Oscillation V-t and (f) FFT curves for different values of normalized optical power. (g) ORN voltage filtered at 28 KHz with different bandwidth and (h) filtered at different center frequencies with a bandwidth of 200 Hz as a function of incident optical power. (i) Analytical approximation of simulated behavior.

Supplementary Figure S8a shows the schematic of an ORN circuit. The inductor and the intrinsic capacitance of the photodetector creates an LC tank circuit that can be forced into relaxation

oscillation by the NDR characteristic of the photodetector. The following differential equation governs the oscillation behavior of the circuit.

$$C\frac{d^2V_{osc}}{dt^2} + G(V_{osc}, P)\frac{dV_{osc}}{dt} + \frac{V_{osc}}{L} = \frac{V_{applied}}{L}$$

$$\sqrt{LC}\frac{d^2V_{osc}}{dt^2} + \frac{P}{P_0}\sqrt{\frac{L}{C}}G(V_{osc}, P_0)\frac{dV_{osc}}{dt} + \frac{V_{osc}}{\sqrt{LC}} = \frac{V_{applied}}{\sqrt{LC}}$$

Here, $G(V_{osc}, P)$ is the negative differential conductance (NDC) of the photodetector which is nonlinear to V_{osc} and can be approximated linear to P via this relation, $G(V_{osc}, P) = \frac{P}{P_0}\sqrt{\frac{L}{C}}G(V_{osc}, P_0)$. This linear dependence on P holds as long as $P \geq P_{NDR,\,min}$ where $P_{NDR,\,min}$ is the smallest optical power that allows the device to exhibit NDR behavior. It is noteworthy that choice of P_0 is arbitrary since it does not affect the absolute value of $G(V_{osc}, P)$. The equation resembles that of a generalized Van der Pol relaxation oscillator. Such a nonlinear differential equation does not have any analytical solution. For an SGM photodetector, $G(V_{osc}, P_0)$ and C are both known quantities as obtained from experimental measurements. An external inductor (L), the incident optical power (P), and the applied voltage ($V_{applied}$) then determine the oscillation behavior. Supplementary Figure S8b shows the colormap of power generated ($P_{gen} = \frac{1}{T}\int_0^T V_{applied} i_L dt$) at the voltage source, $V_{applied}$, due to the photovoltaic regime of operation as a function of $V_{applied}$ and P (assuming C = 1 nF and L = 100 mH). Since the photodetector shows NDR behavior only within -100 mV and 80 mV, the circuit does not show any oscillation at $V_{applied} < -100$ mV and $V_{applied} > 80$ mV. There is a net generation of electrical power ($P_{gen} > 0$) at -100 mV < $V_{applied}$ < 0 mV and a net consumption ($P_{gen} < 0$) at 0 mV < $V_{applied}$ < 80 mV. This is a unique characteristic of SGM

photodetector which separates it from other photoactivated NDR devices where absence of the open circuit voltage prevents any electrical power generation from optical power.

While the equation can be solved for individual values of P, L, and C, it is of great benefit to find a normalized solution to the equation in order to better understand the oscillation behavior of the ORN circuit. A normalized quantity $\frac{P}{P_0}\sqrt{\frac{L}{C}}$ absorbs the change in P, L, C and therefore leads to a normalized solution. Supplementary Figure S8c shows the colormap of fundamental oscillation frequency ($f_{peak,1}$) as a function of $V_{applied}$ and $\frac{P}{P_0}\sqrt{\frac{L}{C}}$ where we have considered P_0 = 1 mW. For $V_{applied}$ = 0V, Supplementary Figure S8d shows the change in $f_{peak,1}$ and oscillation amplitude as a function of $\frac{P}{P_0}\sqrt{\frac{L}{C}}$. With increasing $\frac{P}{P_0}\sqrt{\frac{L}{C}}$, we observe a decrease in $f_{peak,1}$ from its pristine harmonic oscillation frequency of $f_{LC} = \frac{1}{2\pi\sqrt{LC}}$. This decrease can approximately be modeled as $f_{peak,1} = f_{LC}\left(1 - y\frac{P}{P_0}\sqrt{\frac{L}{C}}\right) = f_{LC}(1 - xP)$. For $\frac{P}{P_0}\sqrt{\frac{L}{C}} \approx 400\ \Omega$, there is a drastic increase in oscillation amplitude marking the onset of the relaxation oscillation. This normalized threshold for oscillation helps us predict the oscillation behavior for extreme scaling of device area or optical power. When the device is scaled down to a smaller area, there is a decrease in both P and C for the same optical intensity. Such a change can be compensated directly by increasing L if $P \geq P_{NDR,\,min}$. Similarly, when a device operates at a smaller optical intensity and hence a smaller P ($\geq P_{NDR,\,min}$) for the same device area, L can be increased accordingly to ascertain oscillation from the circuit. It is also important to note that these equations do not consider the effect of parasitic series resistance in the circuit. When P increases significantly so that the series resistance is smaller compared to the resistance of the photodetector, the resistive losses in the circuit may prevent the oscillation from starting. Therefore, the dynamic range for an ORN is determined by the lowest power at which we

observe NDR ($P_{NDR,\ min}$) and the highest power at which the losses from the series resistance become too large.

The origin of these results can be understood through analytical modeling of the oscillation frequency spectra of an ORN. As observed from supplementary Figure S8f, frequency spectrum of a single ORN can be expressed as a sum of all the frequency peaks, including the fundamental and the harmonics. We can model these peaks as Lorentzian peaks centered at $f_{peak,i}$ and of width Δf_i.

$$V_{osc}(P,f) = \sum\nolimits_i \frac{\beta_i}{(f - f_{peak,i})^2 + (\Delta f_i)^2}$$

A bandpass filtering operation at center frequency f with bandwidth BW gives us the integrated amplitude contribution.

$$V_{osc}(P,f,BW) = \int_{f-BW/2}^{f+BW/2} \sum\nolimits_i \frac{\beta_i}{(f - f_{peak,i})^2 + (\Delta f_i)^2} df$$

$$V_{osc}(P,f,BW) = \sum\nolimits_i \int_{f-BW/2}^{f+BW/2} \frac{\beta_i}{(f - f_{peak,i})^2 + (\Delta f_i)^2}\ df$$

$$V_{osc}(P,f,BW) = \sum\nolimits_i \frac{\beta_i}{\Delta f_i}\left(tan^{-1}\left(\frac{f + BW/2 - f_{peak,i}}{\Delta f_i}\right) - tan^{-1}\left(\frac{f - BW/2 - f_{peak,i}}{\Delta f_i}\right)\right)$$

$$V_{osc}(P,f,BW) = \sum\nolimits_i \frac{\beta_i}{\Delta f_i} tan^{-1}\left(\frac{BW.\Delta f_i}{(\Delta f_i)^2 + (f - f_{peak,i})^2 + (BW)^2/4}\right)$$

For small arguments of arctan function,

$$V_{osc}(P,f,BW) \approx \sum\nolimits_i \frac{BW.\beta_i}{(\Delta f_i)^2 + (f - f_{peak,i})^2 + (BW)^2/4}$$

$$V_{osc}(P,f,BW) = \sum_i \frac{BW.\beta_i}{(\Delta f_i)^2 + (f - i.f_{LC}(1 - xP))^2 + (BW)^2/4}$$

$$V_{osc}(P,f,BW) = \sum_i \frac{BW.\beta_i}{(\Delta f_i)^2 + (i.x.f_{LC})^2 \left(P - \left(\frac{f_{LC} - f}{i.x.f_{LC}}\right)\right)^2 + (BW)^2/4}$$

$$V_{osc}(P,f,BW) = \sum_i \frac{\gamma_i}{(\Delta P_i)^2 + (P - P_i)^2}$$

Here, $\gamma_i = \frac{BW.\beta_i}{(i.x.f_{LC})^2}$, $\Delta P_i = \frac{(\Delta f_i)^2 + \frac{(BW)^2}{4}}{(i.x.f_{LC})^2}$, and $P_i = \frac{f_{LC} - f}{i.x.f_{LC}}$. The final equation shows that the amplitude contribution at each center frequency is also a sum of Lorentzian peaks. In addition, increasing filter bandwidth causes an increase in the peak width (ΔP_i) and increasing the center frequency causes a shift in these peaks (P_i) towards smaller optical powers.

S3. ORN kernel experimental V-t curves for different intensities

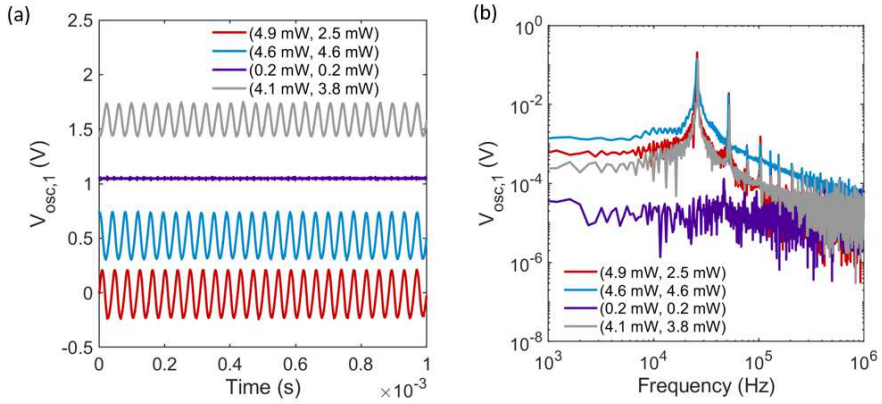


Figure S9: Representative oscillation time series and corresponding frequency spectra for 1x2 ORN kernel.

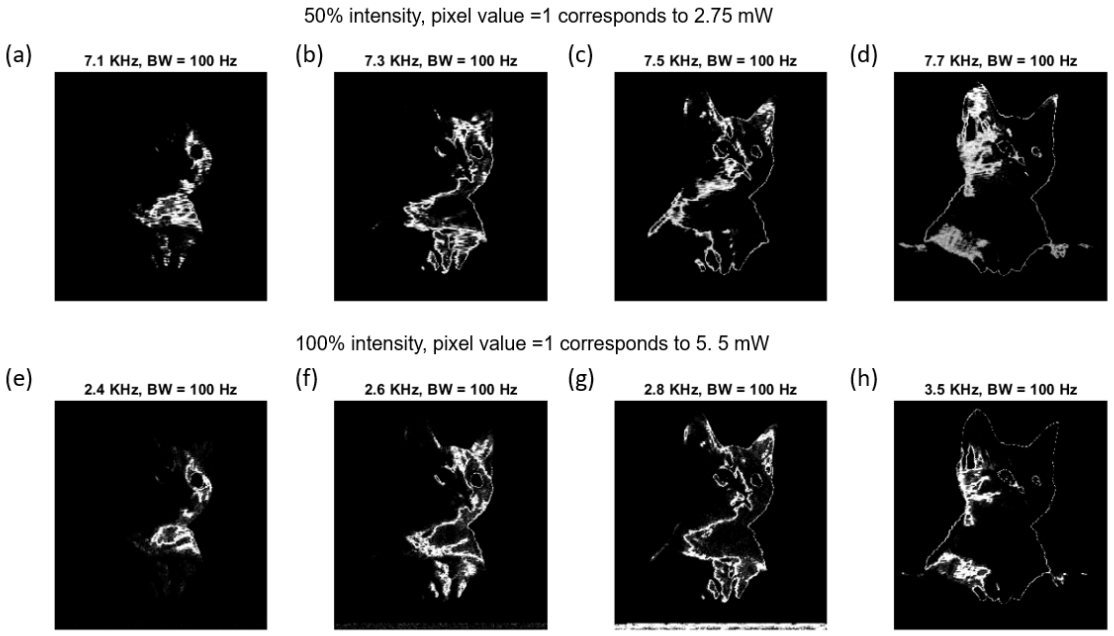


Figure S10: Experimental results for images processed at different frequencies by the cascaded circuit when pixel value = 1 corresponds to an optical power of (a-d) 2.75 mW and (e-h) 5.5 mW.

S4. Processed images for a 10x10 kernel with nearest neighbor coupling

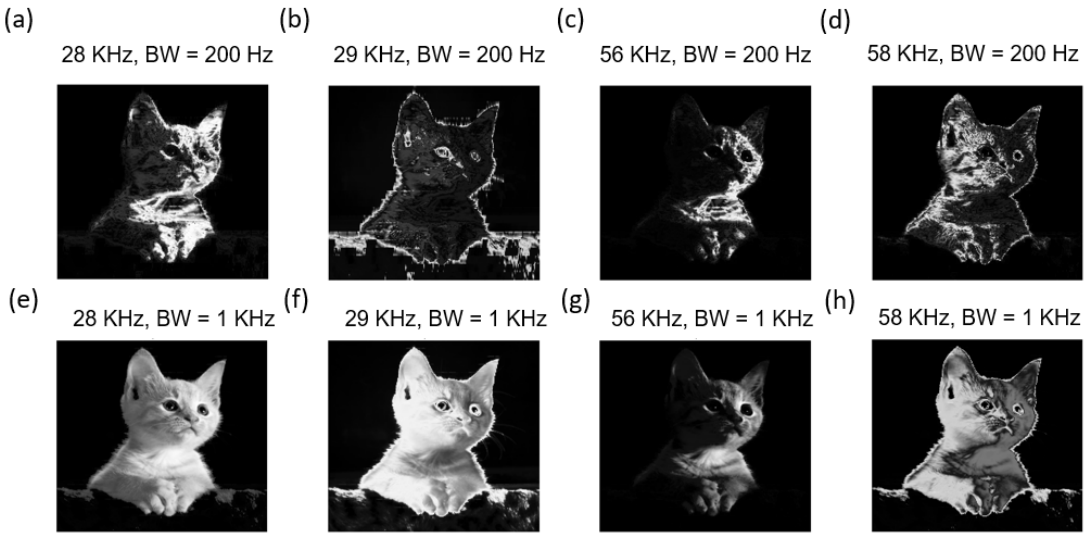


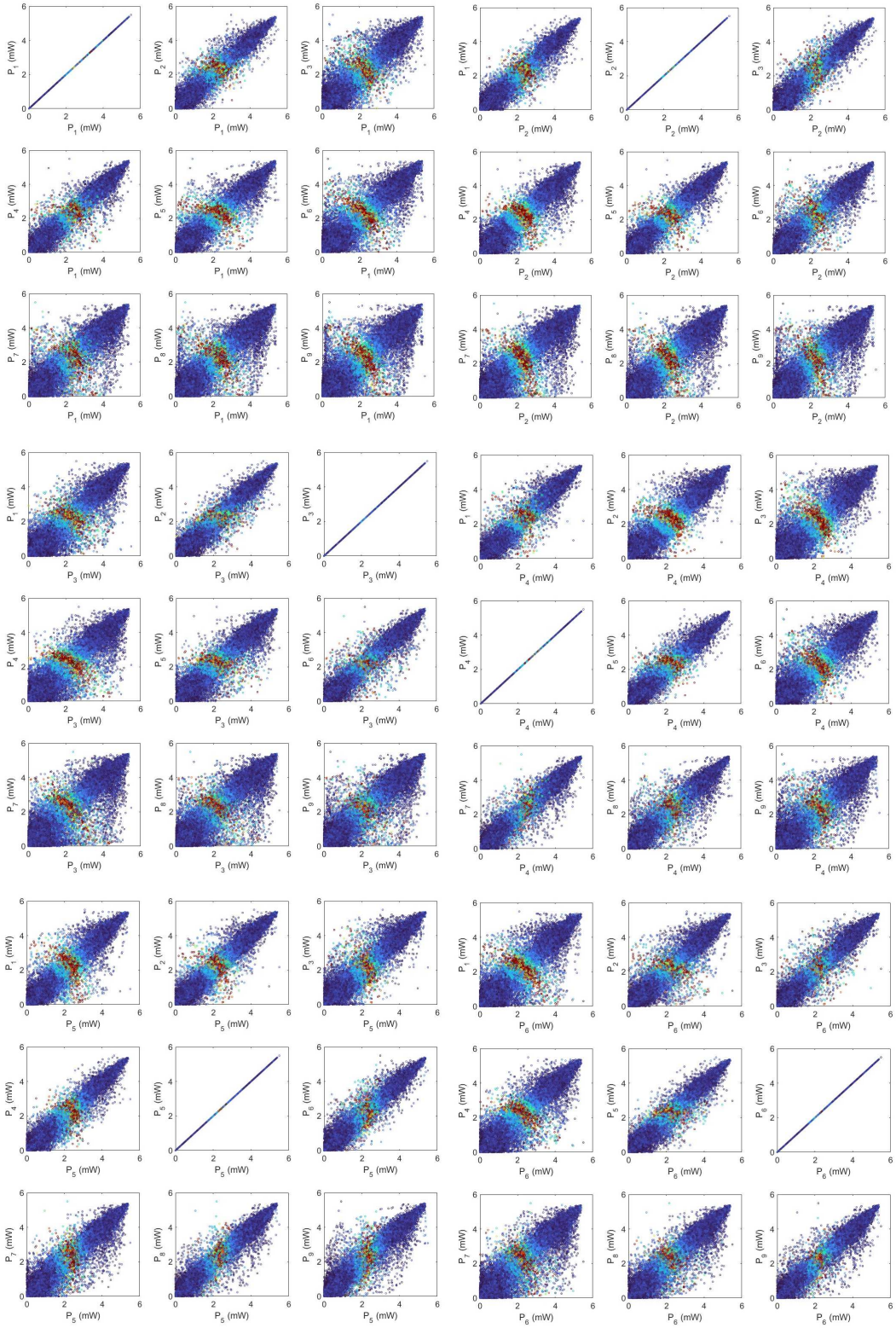
Figure S11: Simulation results for a 10x10 kernel where only nearest neighbor oscillators are coupled by an inductance of 5H

S5. Functions implemented by 3x3 ORN kernel

Since the 3x3 ORN kernel takes 9 optical powers (P_1, P_2, …, P_9) as inputs, it has a 9D input space. For any image, a pixel intensity range of 0 to 1 can be mapped to an optical power range of 0 to P_{max}. If this range is divided into N linearly spaced points, the total number of input combinations become N^9. For 1x2 kernel case, we considered N = 50 and therefore simulated for $50^2 = 2500$ different combinations of P_1 and P_2. However, if we were to do the same for 9 oscillators, the number of combinations become $50^9 \approx 2\text{x}10^{15}$ which would require an impractically large amount of time and data storage. As a result, we have adopted a different approach to understanding the mathematical functions implemented by the circuit at different frequencies. Rather than simulating over the whole possible input space, we only look at the outputs returned by the input combinations encountered for the original cat image. Since 9D inputs cannot be visualized the same way as 2D inputs, we plot the output as colormaps for all 2D combinations of P_i and P_j where $1 \leq i, j \leq 9$ for a frequency band centered at 4 KHz with a bandwidth of 100 Hz as shown in supplementary Figures S12a-i. All the colormaps show output voltage this frequency band within a range of 0 to 0.2V. According to these results, when P_i and P_j average to a value of ~2.5 mW, we see a peak output voltage for the given band. If it were a linear average, i.e., $P_{av} = \frac{P_i + P_j}{2}$, we would observe this peak surface as a straight line. However, we see a considerable curvature more similar to that of a circle. As shown in supplementary Figure S13a (lower panel), when we plot the output voltage as a function of the root mean square of all P_i ($P_{RMS} = \sqrt{\frac{1}{9}\Sigma_{i=1}^{9} P_i^2}$), we see a peak around P_{RMS} = 2.5 mW with a peak width of ~1 mW. It is important to note that the choice of RMS average stems from the graphical observation of circular peak surface shape and it is an approximation. This function essentially causes a subspace of all possible inputs, inputs that average to ~2.5 mW optical

power, to return higher voltages while returning lower voltages for all other combinations. Supplementary Figure S13a (upper panel) shows the corresponding processed image due to this function. This image predominantly detects the edges of the cat. Supplementary Figure S13b (upper panel) shows the output voltage as a function of $P_{diff,RMS}(=\sqrt{\frac{1}{9}\Sigma_{i=1}^{9}(P_i - P_5)^2})$. When $P_{diff,RMS}$ is small, the pixels that return high voltage are not true edges of the image. On the other hand, we get the true edges when $P_{diff,RMS}$ is large. As a result, the processed image is essentially a combination of both of a background (small $P_{diff,RMS}$) and the edges (large $P_{diff,RMS}$) as shown in supplementary Figure S13b (lower panel). For this frequency band at 4 KHz (200 Hz bandwidth), the average optical power for peak voltage falls at the 2.5 mW range where the background pixels are less prominent for this cat image. At different frequency bands, the ORN kernel therefore will detect edges of different ranges which will come with their own "background." However, for any given image, there will be a certain range where such backgrounds are considerably small and we will be able to get a good edge detection. Such an edge detection method is not universally applicable since different input images will show better edge detection at different frequency bands due to the differences in input pixel intensity ranges. On the other hand, a convolution based edge

detection operator, such as Sobel operator, will detect edges at all intensity ranges and such a background will not be present there. While this lack of background is advantageous for Sobel



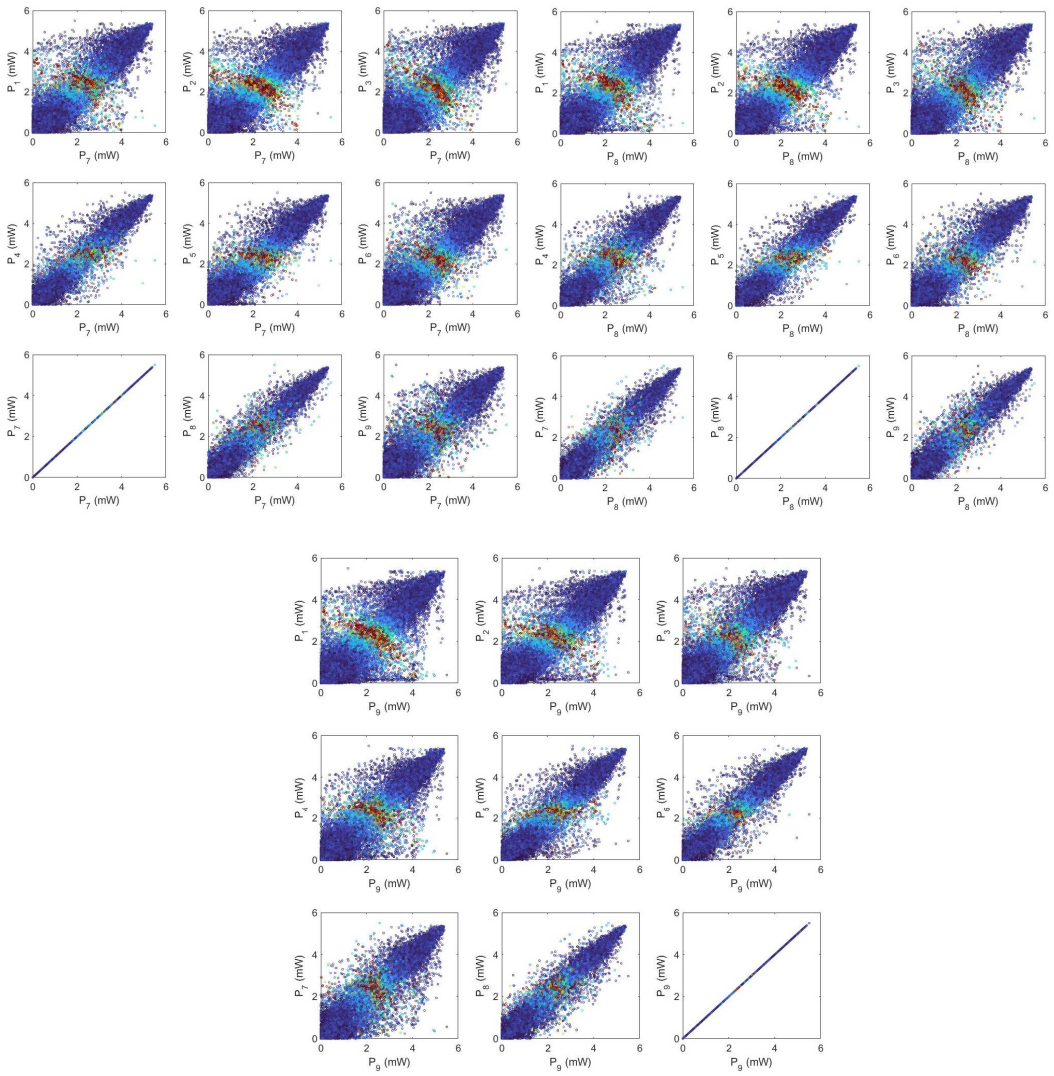


Figure S12: Output voltage colormaps for different combinations of $P_1, P_2, \ldots, P_9$ for the input image.

operator, it will also return many insignificant edges which then need to be filtered out. It is difficult to comment on which method is better since definition of edges depends heavily on application. While more detailed edge detection may be preferred in one application, edges outlining the object may be more important to another application.

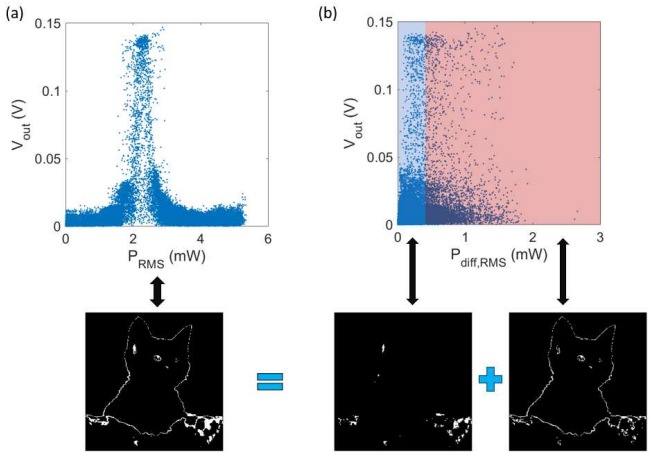


Figure S13: Output voltage as a function of (a) RMS optical power of all pixels and (b) RMS optical power difference with the central pixel and corresponding processed images.

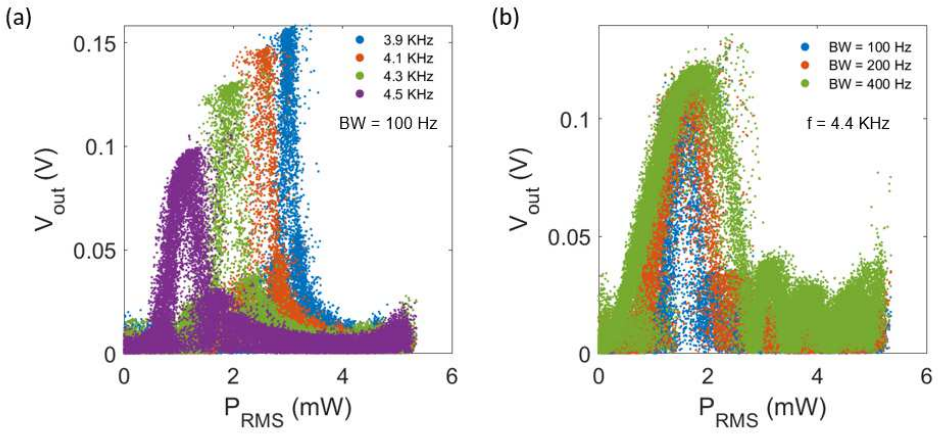


Figure S14: Output voltage as a function of P_{RMS} for (a) different center frequencies and (b) bandwidths.

Supplementary Figure S14 shows the output voltages as a function of P_{RMS} for different frequency bands for different bandwidths. Similar to the single oscillator case, a larger bandwidth widens the peak. Therefore, a frequency band with a larger bandwidth will return higher voltages for more combinations of input optical powers, allowing a broader range of optical powers to be grouped together. When a number of continuous pixels are within a similar intensity range, they all return similar output as if they form the same object, leading to an object segmentation operation. This

is similar to Otsu's thresholding approach for object segmentation where the pixel intensities are thresholded at different intensity ranges to identify the objects within a picture.

S6. Readout layer:

A three-layer simple neural network is constructed for the training and testing. The input layer contains 49×k neurons where k is the number of frequency samples considered. The hidden layer consists of 100 neurons with ReLU activation function and output layer consists of 10 neurons, each of which indicates the possibility of corresponding classes. The three layers here are fully connected, suggesting 49000×k synapses or weight values $\boldsymbol{w}$ need to be trained for good classification. In each training epoch, 8571 input images are used as training images, the other 1429 are used as test images. In the supervised training process, one-hot encoding and the softmax activation function : $\phi_m(\boldsymbol{I}) = {e^{I_m\xi}}\Big/{\sum_{k=1}^{M} e^{I_k\xi}}$ are used, here ξ is a scaling factor. And the loss (or cost) that need to be minimized is calculated by the cross-entropy function: $Loss = -\frac{1}{M}\sum_{m=1}^{M} y_m \log\left[\phi_m(\boldsymbol{I})\right]$, where y_m is the label of the input image, $M = 10$ is the number of classes. The values of the weight are initially randomly generated from a Gaussian distribution, then updated by gradient descent and backpropagation: $\boldsymbol{w} = \boldsymbol{w} - \eta\frac{\partial Loss}{\partial \boldsymbol{w}}$, here the learning rate $\eta = 0.2$. In images testing, the most activated neuron (with the highest output value) is compared with the label, and the accuracy is calculated by dividing the matched images number with the testing image number.

S7. LSM details

S7.1 LSM experiments and data handling:

These experiments use an inductively cascaded array of 9 ORNs as the liquid layer. An MNIST image is typically 28×28 pixels. We truncate the image to 21×21 pixels and then divide it into a 7×7 cell matrix of 3×3 pixels. This is similar to sweeping the image with a 3×3 convolutional kernel with a stride of 3. Then each of these 3×3 pixels cell is physically projected onto the liquid layer. Oscillation V-t time series is then collected from one of the pixels (7th oscillator of the array in these experiments) using an oscilloscope. The acquired time series is then transformed to a frequency spectrum via FFT. The FFT operation is equivalent to performing a bandpass filtering operation at different frequencies. For these measurements, we chose 1 KHz bandwidth for the bandpass filter and performed this operation for 100 different frequency bands up to 100 KHz. Therefore, we get a 7×7 data for every frequency sample, a 7×7×100 data per image, and a total dataset of size 7×7×100×10000. Then 6/7 part (~86%) of the data was separated as the training dataset (7×7×100×8571) and the rest 1/7 part (~14%) was separated as the testing dataset (7×7×100×1429). We have then trained the readout network using 100 different 7×7×8571 datasets and tested them with the corresponding testing datasets.

S7.2 Simulation of the LSM:

Simulation of the LSM requires us to solve the equations governing the dynamics of all oscillators in the ORN kernel simultaneously. Let V_k be the instantaneous voltage across the k^{th} SGM photodetector device. Then V_k is updated at each time step using the following equations:

$$V_k(t + \Delta t) = V_k(t) + \frac{dV_k}{dt}(t) \times \Delta t$$

$$i_{L,k}(t + \Delta t) = i_{L,k}(t) + \frac{di_{L,k}}{dt}(t) \times \Delta t$$

$$\frac{dV_k}{dt}(t) = \frac{i_{c,k}(t)}{C_k}$$

$$\frac{di_{L,k}}{dt}(t) = \frac{V_k(t) - V_{k-1}(t)}{L_k}$$

$$i_{c,k}(t) + i_{d,k}(t) + i_{L,k}(t) = i_{L,k+1}(t)$$

$$i_{d,k}(t) = I(V_k(t))$$

Here, $i_{L,k}$, $i_{c,k}$, $i_{d,k}$ are the currents through the kth inductor, capacitor, and photodetector. $i_{d,k}$ is obtained from a lookup table containing the experimental I-V data for our devices. In addition, $V_k(0) = 0V$ and $i_L(0) = 0A$ for all k, and $V_0(t) = 0V$ and $i_{L,10}(t) = 0A$ for all t. In order to make sure that the solution is not diverging, we need to consider small enough Δt appropriate for the problem at hand. Supplementary Figure S15 shows the circuit schematic for this kernel. Every time a new 3×3 pattern is introduced to the kernel, this system of equations is solved simultaneously, and the corresponding time series and frequency spectrum is stored.

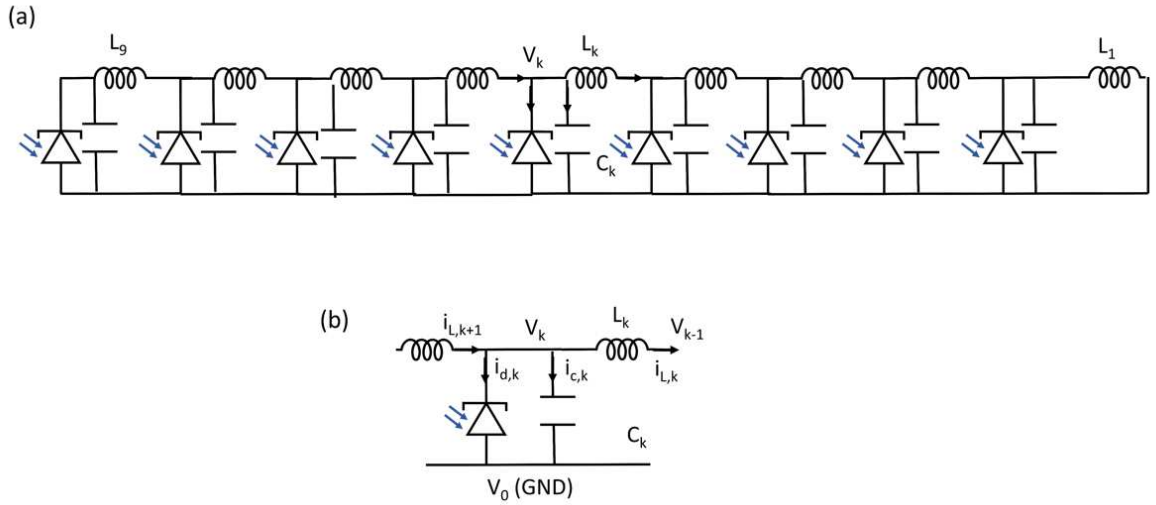


Figure S15: (a)Circuit schematic of ORN kernel and (b) current and voltage components

S8. Energy calculation for peripheral circuit

An ORN array requires peripheral circuitry to read the oscillating voltage signals and filter them into different frequency bands. Power (voltage) in each frequency band mapped across the ORN array then represents a single feature of the image. There are 3 key components to the peripheral circuit: (1) amplifiers to isolate the oscillators from the readout circuit and provide necessary electrical energy to sampler (2) sample and hold circuit consisting of sampling switches and capacitors and (3) a DFT (discrete Fourier transform) processor that filters the input signal into frequency bands. An on-chip analog DFT processor can be designed using the radix-2 decimation in time (DIT) FFT algorithm. The radix-2 algorithm can be implemented in voltage, current or charge domain. However, the charge domain approach has been the most promising one in terms of energy efficiency of FFT operation. There has been a demonstration of an analog FFT processor fabricated with CMOS 65 nm technology using charge domain approach that can perform a 16 point FFT with an energy cost of 12.2 pJ at a sampling speed of 5 GHz[1]. We can

scale this approach for 128-point FFT so that the ORN array oscillating signals can be filtered into 64 unique frequency bands and therefore generates 64 different features per oscillator.

A k-point implementation of radix-2 algorithm has $\frac{1}{2}klog_2k$ butterfly operations and each butterfly operation requires 10 scalar operations. Therefore, the total number of operations in a k-point FFT is $5klog_2k$. Scaling a 16-point FFT architecture to a 128-point architecture therefore increases the scalar operations by 14x. If we use the same design architecture as Ref X, the number of transistor switches and total switching energy consumption in the processor would then increase by 14x. On the other hand, the total number of sampling capacitors increases linearly from 256 to 2048. It is noteworthy that the destructive nature of passive computing requires multiple copies of the same sample. This implementation requires 2 copies per butterfly, 2 copies for complex math, 2 copies for differential input, and 2 copies for quadrature (I/Q) input which makes the total number of capacitors 2×2×2×2=16 times larger than the FFT points. However, since our signals are real valued, we can multiplex signals from two oscillators into the same FFT architecture which effectively reduces the number of sampling capacitors and switches by 2x per oscillator. In addition, since the quadrature inputs are not available for our signals, we can reduce the number of sampling capacitors and switches by another 2x per oscillator. However, this also reduces the number of useful frequency bins by 2x. Therefore, the 128-point FFT would have 512 sampling capacitors and have a 3.5x larger energy consumption compared to 16 point FFT and bin the signal to 64 frequency bands.

The choice of sampling capacitor directly depends on the maximum SNR we expect at full scale (V_{FS}). Our ORN arrays typically have a maximum rms voltage of 0.3V (= V_{FS}) and the output reflected noise power contribution for this FFT architecture is ~$0.26\frac{kT}{C}$ where C is the value of the

sampling capacitance. In order to compare our results to that of a digital processor, we choose an SNR of 49.92 dB which is equivalent to the SNR of an 8 bit ADC ($b = \frac{SNR-1.76}{6.02}$). Then the sampling capacitance C is given by

$$C = 10^{\frac{SNR}{10}} \times \frac{0.26kT}{{V_{FS}}^2} = 0.3976 \times 4^b \times \frac{k_B T}{V_{FS}^2} = 1.2\ fF$$

Since the energy consumption of 16 point FFT processor is 12.2 pJ/FFT, the total capacitance of the switching transistors is $\frac{2\times 12.2\times 10^{-12}}{1.2^2} = 17\ pF$ where $V_{DD} = 1.2V$ is the supply voltage for the transistors. The number of transistors for 16 point FFT is ~5120 and therefore the gate capacitance of each transistor is 3.31 fF. However, the on resistance of the transistors is given by $\sim\frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS}-V_{th})}$ and therefore the maximum switching frequency is $\sim\frac{\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS}-V_{th})}{7.824C}$. Numerical calculations (supplementary Figure S16) show that the capacitor voltage can settle to 2% of the final voltage within the half cycle of 1 GHz clock for a $V_{GS} - V_{th} = 0.16V$ even for the worst case scenario where the voltage rises from 0V to $V_{FS} = 0.3V$. Since $V_{th} \approx 0.3V$ for NMOS switches, we can choose a reduced $V_{DD} = 0.46V$ with a switching transistor gate capacitance of $C_{SW} = C_{ox}A = 0.05$ fF for W = L = 65 nm. Here, we will work out the energy consumption for an example of 128-point FFT processor. For 128-point FFT, the number of transistors would be 3.5×5120 = 17920 and the total switching energy consumption will be $17920 \times \frac{1}{2} \times 5 \times 10^{-17} \times 0.46^2$ J = 47.17 fJ.

On the other hand, each signal needs to be sampled 128 times using a pseudo-differential sampler that has 4 transistors. For the same C_{SW}, the switching energy for the sampler is $4 \times 128 \times \frac{1}{2} \times 5 \times 10^{-17} \times 0.46^2$ J = 2.7 fJ.

The sampling capacitors in the FFT processor have to be charged by the oscillator output and this charging costs an energy of $4 \times 128 \times \frac{1}{2} \times 1.2 \times 10^{-15} \times 0.15^2$ J = 6.9 fJ. Since the oscillators can have any instantaneous voltage between 0 and 0.3V, an averaged 0.15V was considered for this calculation.

This brings the total energy consumption of the peripheral circuit to 56.8 fJ. Since the FFT processor generates 64 outputs and each convolution operation for a 3x3 kernel is equivalent to 17 operations, the energy cost becomes 56.8/(64x17) fJ/OP= 0.0522 fJ/OP and the processor performance is 1/0.0522 Ops/J = 19159 TOPS/W. Supplementary Figure S16 shows the processor performance and switching speed as a function of V_{DD} in the switching MOSFETs. At a reduced operation frequency of 1 MHz, the projected performance increases to 42211 TOPS/W.

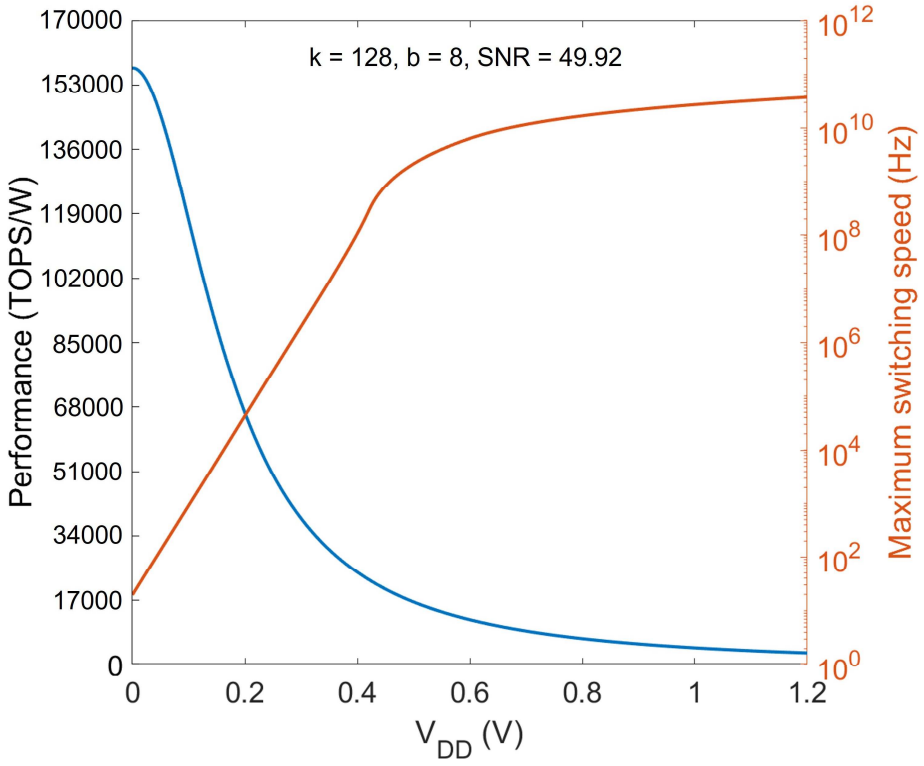


Figure S16: Performance and maximum switching speed as a function of MOSFET V_{DD}.

Extended Table S1 shows the performance comparison between different NPUs and general purpose neural engines.

NPU application	**Type**	**Comment**	**bits**	**Reported TOPS/W**	**Normalized TOPS/W (8 bits)**
VMM[2]	Digital	NN accelerator	8	1.7	1.7
MAC macro[3]	Digital	Mobile SoC	8	11.5	11.5
VMM[4]	Analog	RRAM	6	9	5.06
VMM[5]	Digital	3D-RRAM	8	8.32	8.32
VMM[6]	Digital	RRAM	4	36.61	9.15
VMM[7]	Analog	Magnetic RRAM	1	405	6.33
Logic gate[8]	Digital	2D neuristor	1	622.35	9.72
Analog to information conversion[9]	Analog	In-sensor NN	8	43.5	43.5

VMM[10]	Digital	SRAM	4	351	87.75
VMM[11]	Analog	RRAM	1	78.4	1.23
MAC macro[12]	Analog	DRAM	4	217	54.25
VMM[13]	Digital	SRAM	1	885.86	13.84
VMM[14]	Digital	BNN accelerator	1	223	3.48
ReRAM write[15]	Analog	RRAM	4	66.5	16.63
VMM[16]	Digital	DNN learning processor	8	146.52	146.52
Arithmetic logic[17]	Digital	Superconducting logic devices	8	120	120
VMM[18]	Mixed	NN accelerator	4	161.6	40.4
VMM[19]	Analog	Si-CMOS/CAAC-IGZO based memory	6	210	118.13
VMM[20]	Digital	Stochastic NN accelerator	8	75	75
VMM[21]	Digital	BNN accelerator	1	617	9.64
VMM[22]	Digital	BNN accelerator	2	198.9	12.43
SNN[23]	Mixed	Mixed mode spiking neuron	4	124.2	31.05
Combinatorial problem solving[24]	Analog	Binary RRAM	1	2000	31.25
MAC macro[25]	Digital	SRAM	8	63	63
MAC macro[26]	Analog	SONOS memory	8	100	100
MAC macro[27]	Analog	SRAM	4	177	44.25
MAC macro[28]	Digital	SRAM+DRAM	1	1220	19.06
Image denoising [29]	Digital	SRAM	1	51.3	0.8
Image filtering[30]	Digital	SRAM	1	389	6.08
MAC macro[31]	Digital	SRAM	8	30.3	30.3
Data buffering macro[32]	Digital	SRAM	4	43	10.75
MAC macro[33]	Digital	SRAM	1	20943	327.23
VMM[34]	Analog	Ferroelectric tunnel junction memristors	4	100	25
General purpose	Digital	NVIDIA A100	8	4.992	4.992

General purpose	Digital	Google TPUv4	8	1.432	1.432
General purpose	Digital	Apple a16 Bionic	8	2.67	2.67
General purpose	Digital	Huawei Da Vinci Ascend 310	8	2	2
General purpose	Digital	Qualcomm Snapdragon 865	8	4.5	4.5
Nonlinear convolution (3x3 kernel) (This work)	Analog	ORN	8	50-67000	50-67000

References

1 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).

2 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).

3 Song, J. *et al.* in *2019 IEEE International Solid-State Circuits Conference-(ISSCC).* 130-132 (IEEE).

4 Wan, W. *et al.* A compute-in-memory chip based on resistive random-access memory. *Nature* **608**, 504-512 (2022).

5 Huo, Q. *et al.* A computing-in-memory macro based on three-dimensional resistive random-access memory. *Nature Electronics* **5**, 469-477 (2022).

6 Xue, C.-X. *et al.* A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices. *Nature Electronics* **4**, 81-90 (2021).

7 Jung, S. *et al.* A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601**, 211-216 (2022).

8 Chen, H. *et al.* Logic gates based on neuristors made from two-dimensional materials. *Nature Electronics* **4**, 399-404 (2021).

9 Sadasivuni, S., Bhanushali, S. P., Banerjee, I. & Sanyal, A. In-sensor neural network for high energy efficiency analog-to-information conversion. *Scientific reports* **12**, 18253 (2022).

10 Dong, Q. *et al.* in *2020 IEEE International Solid-State Circuits Conference-(ISSCC).* 242-244 (IEEE).

11 Liu, Q. *et al.* in *2020 ieee international solid-state circuits conference-(isscc).* 500-502 (IEEE).

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

13 Kuo, S.-H. & Chang, T.-S. in *2022 IEEE International Symposium on Circuits and Systems (ISCAS).* 2968-2972 (IEEE).

14 Andri, R., Karunaratne, G., Cavigelli, L. & Benini, L. in *2021 IEEE International Symposium on Circuits and Systems (ISCAS).* 1-5 (IEEE).

15 Mochida, R. *et al.* in *2018 IEEE Symposium on VLSI Technology.* 175-176 (IEEE).

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

17 Nagaoka, I. *et al.* in *2019 IEEE International Superconductive Electronics Conference (ISEC).* 1-3 (IEEE).
18 Jo, W. *et al.* in *2022 IEEE International Symposium on Circuits and Systems (ISCAS).* 365-369 (IEEE).
19 Chen, M.-C. *et al.* in *2022 International Electron Devices Meeting (IEDM).* 18.12. 11-18.12. 14 (IEEE).
20 Romaszkan, W. *et al.* A 4.4–75-TOPS/W 14-nm Programmable, Performance-and Precision-Tunable All-Digital Stochastic Computing Neural Network Inference Accelerator. *IEEE Solid-State Circuits Letters* **5**, 206-209 (2022).
21 Knag, P. C. *et al.* A 617-TOPS/W all-digital binary neural network accelerator in 10-nm FinFET CMOS. *IEEE journal of solid-state circuits* **56**, 1082-1092 (2020).
22 Hu, Y. *et al.* A 28-nm 198.9-TOPS/W Fault-Tolerant Stochastic Computing Neural Network Processor. *IEEE Solid-State Circuits Letters* **5**, 198-201 (2022).
23 Kim, S. *et al.* in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits).* 38-39 (IEEE).
24 Hong, M.-C. *et al.* in *2021 IEEE International Electron Devices Meeting (IEDM).* 21.23. 21-21.23. 24 (IEEE).
25 Fujiwara, H. *et al.* in *2022 IEEE International Solid-State Circuits Conference (ISSCC).* 1-3 (IEEE).
26 Agrawal, V. *et al.* in *2020 IEEE International Memory Workshop (IMW).* 1-4 (IEEE).
27 Zhang, B. *et al.* in *2022 IEEE Custom Integrated Circuits Conference (CICC).* 1-2 (IEEE).
28 Zhang, X. & Basu, A. in *2022 IEEE Custom Integrated Circuits Conference (CICC).* 1-2 (IEEE).
29 Bose, S. K., Singla, D. & Basu, A. A 51.3-TOPS/W, 134.4-GOPS In-Memory Binary Image Filtering in 65-nm CMOS. *IEEE Journal of Solid-State Circuits* **57**, 323-335 (2021).
30 Bose, S. K. & Basu, A. in *2021 IEEE Asian Solid-State Circuits Conference (A-SSCC).* 1-3 (IEEE).
31 Lee, C.-F. *et al.* in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits).* 24-25 (IEEE).
32 Rajanna, V. K., Taneja, S. & Alioto, M. in *ESSCIRC 2021-IEEE 47th European Solid State Circuits Conference (ESSCIRC).* 127-130 (IEEE).
33 Lin, C.-S. *et al.* in *2021 IEEE Asian Solid-State Circuits Conference (A-SSCC).* 1-3 (IEEE).
34 Berdan, R. *et al.* Low-power linear computation using nonlinear ferroelectric tunnel junction memristors. *Nature Electronics* **3**, 259-266 (2020).