1 TOPS optical neuromorphic processing with Kerr soliton crystal microcombs: Scaling the network in size and speed

David J Moss

11 TOPS optical neuromorphic processing with Kerr soliton crystal microcombs: Scaling the network in size and speed

David J. Moss[1]

[1] Optical Sciences Centre, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

ABSTRACT

Artificial neural networks (ANNs) can extract the hierarchical features of raw data and are of significant interest for machine learning tasks such as computer vision, speech recognition, playing board games and medical diagnosis [1-7]. Optical neural networks in turn have the capability to dramatically accelerate the computing speed of ANNs to overcome the intrinsic bottleneck in bandwidth that electronics is subject to. Convolutional neural networks (CNNs), are inspired by biological systems such as the visual cortex, and are a powerful approach to greatly reduce the parametric network complexity in order to enhance the accuracy of the predictions of the system. In this paper, we demonstrate a universal optical convolutional accelerator that can be used in conjunction with both electronic and optical neural networks. It operates beyond 10 Tera-OPS (TOPS - operations per second) and produces convolutions of extremely large scale images of 250,000 pixels in size with a resolution of 8-bits. It generates 10 convolutions simultaneously in parallel, with 10 different kernels. This processing simultaneously — enough for facial image recognition. After demonstrating this, we then use the exact hardware to form a convolutional neural network consisting of a convolutional front-end followed by a deep optical neural network fully connected layer, together forming a CNN with ten neurons at the output. We successfully perform the recognition of all 10 hand written digits, each consisting of 900 pixel handwritten digit images. We achieve an accuracy of 88% which is very close to the theoretical accuracy of 90%. Our results are based on simultaneously interleaving temporal, wavelength and spatial dimensions enabled by an integrated microcomb source. We present a performance comparison of the different optical neural networks that have been published. Finally, we explicitly demonstrate that our method is intrinsically scalable in both size and speed, up to the PetaOPs per second (POPs) regime in speed and to well over 24,000 synapses in size. We perform theoretical evaluation of the scaled system performance and show that it is trainable to much more complex networks for real-world demanding applications including real-time video recognition and autonomous unmanned vehicle control.

Keywords: Optical neural networks, neuromorphic processor, microcomb, convolutional accelerator

1. INTRODUCTION

Artificial neural networks (ANNs) can perform highly complex operations for decision making in applications such as medical diagnosis, playing board games, speech translation, face recognition, and much more [1- 4]. They consist of collections of nodes that have connections that are weighted according to proper feedback in order to adjust the network parameters, so the system can "learn" while it performs. Classical fully connected networks face significant processing challenges, particularly for advanced applications such as extremely high-dimensional data. In order to dramatically simplify the tasks, convolutional neural networks (CNNs) can distill the input data representations from their raw form, in order to predict their properties with greatly simplified parametric complexity as well as highly improved accuracy [5]. CNNs are inspired by the behavior of biological systems such as the visual cortex, and have been widely successful for many applications such as pattern recognition, natural language processing, computer vision, and many other areas [6, 7].

The power of ANNs is determined by the computing capability of the neuromorphic component hardware. In this regard, optical neural networks (ONNs) [8-19] in particular are highly promising approaches towards next-generation neuromorphic systems for computation, since they can overcome the intrinsic speed bottleneck of electronic ANNs [6, 20-23] and so can realize ultra-high speed computing, because of the extremely wide optical bandwidths available, particularly the >10 THz wide optical telecommunications C and L bands [8]. ONNs have attracted huge interest and have formed the basis of major breakthroughs that have been recently reviewed [13-19]. Operating in analog mode, they largely circumvent the limitations inherent in the time and energy taken up during storing and reading data to memory – a well known drawback of classic computer architectures which is known as the von Neumann bottleneck [20]. Significant progress has been achieved in addressing this through the use of high speed, trainable, and highly parallel ONNs [8-19, 24-28]. These are further enhanced by the potential for full integration on a single photonic integrated circuit [8,12,14,15] which would yield computational densities that are extremely high. Nonetheless, despite these breakthroughs there is still significant room for advancement for ONNs. Practical real-world computer vision functions

require the ability to process data at extremely large scales, and this is still extremely difficult because these systems largely contain fully connected networks where the input scale is governed primarily by hardware parallelism. This yields fundamental tradeoffs between the network footprint and its scale. Furthermore, ONNs so far have largely failed to demonstrate the extreme computing speeds that analog photonics can achieve.

Here, we report an optical convolution accelerator that has the ability to process and compress extremely large-scale data and at ultra-high speeds. By interleaving time, wavelength, and spatial dimensions through the novel application of integrated Kerr frequency soliton crystal micro-combs [29 – 143], we realize a vector computing speed of 11.322 Tera-Ops per second (TOPS) and then demonstrate the processing of images that are 250,000 pixels in size, achieving the simultaneous generation of 10 convolution kernels. The convolution accelerator is completely reconfigurable in a fully dynamic sense – without requiring any change in hardware. It is also intrinsically highly scalable. We show that the exact same hardware can operate simultaneously as both a front-end convolutional accelerator generating many parallel kernels simultaneously, as well as performing as a deep optical fully connected convolutional neural network. The CNN is then used for recognition of the full handwritten image set consisting of all ten digits (0-9). We achieve 88% accuracy which is very close to the theoretical accuracy of 90%. Finally, we present detailed architectures that will enable scaling the network both in speed to the Peta-OP per second (POP) regime as well as in size and scale to > 24,000 synapses, which is enough for almost all conceivable applications. We achieve this design through the use of existing commercially available components and equipment based on the S,C,L telecommunications wavelength bands. We theoretically evaluate the scaled systems achieving highly promising results.

Our approach towards optical neural networks is a major breakthrough towards achieving fully monolithically integrated ONNs, and is fundamentally enabled by the integrated Kerr soliton crystal microcomb. Futher, our approach is universal and stand-alone — it is fully compatible with both optical and electrical systems and hence it is able to function as a universal front end for ultrahigh bandwidth data compressing for any type of neuromorphic hardware — either optical or electronic. Our approach will make massive-scale data machine learning for ultrahigh speed data processing in real-time a reality.

2. PRINCIPLE OF OPERATION

Figure 1 illustrates the principle of operation of our optical convolutional accelerator (CA). It contains high-speed electrical input signals and data port outputs. Figure 2 illustrates the experimental configuration. First, the input data vector (X) is encoded serially where the temporal symbol intensity is represented by an electric waveform having a symbol rate of $1/\tau$ (baud), with τ being the symbol period. The convolution kernel is similarly represented by a weight vector W having a length of R, that encodes the optical microcomb power of the lines through the use of spectral shaping with a commercially available system (Waveshaper). Next, the replicas weighted by W are generated by multicasting the time dependent waveform X onto the kernel wavelengths with an electro-optic modulator. Following this, the optical signal is then propagated through a delay, achieved with standard single mode fiber with a dispersion such that the delay step between the adjacent channels wavelength equals the duration of the symbols of X, thus interleaving the wavelength and time dimensions. Finally, the weighted and delayed copies of the signal are detected via high speed photodetection, effectively summing all of the signals in a particular time-slot, with each slot containing the convolution between W and X for each convolution window. Hence, the convolution window then essentially slides along at the baud rate of X, or the modulation speed. Each output symbol is the result of R multiply-and-accumulate (MAC) operations, and the computing speed is then $2R/\tau$ OPS. Since the process speed here scales with the number of wavelengths as well as the baud rate, it can be significantly enhanced to achieve the TOP regime through the use of massively parallel channels of wavelengths generated by a microcomb chip. Importantly, the length of X, the input data, is unlimited. Hence, the convolution accelerator can operate on data at arbitrarily large scales, only limited by the accompanying electronics. Likewise, the number of kernels and their length are also essentially unlimited, subject only to the number of channels or wavelengths. We perform a parallel convolution of many kernels by adding extra sub-bands with each kernel having R wavelengths. After multicasting followed by the dispersive delay, the kernels (sub-bands) are then demultiplexed and separately detected using very high speed photodetectors. This generates a unique and separate electronic waveform, one for every kernel.

Fundamentally, convolutional accelerators process vector data. However, they can in fact perform matrix operations that are needed for processing images, for example, by first effectively flattening the matrix into a 1 dimensional vector. The exact method that this is performed governs the "stride" of the sliding convolutional window as well as the effective computing speed of the matrix operations. The method used to flatten the matrix determines the receptive field (convolution slot) to slide with a horizontal stride of unity, so that every input element of the matrix has a corresponding convolutional output, together with a vertical stride that scales with the convolutional kernel size. A large vertical stride essentially produces sub-sampling of the raw input matrix vertically, which is effectively a partial pooling function [144] on top of the convolution itself. This effectively reduces the matrix computing speed, which is equal to an equivalent overhead cost, that inversely scales with the kernel size. Hence, a 3x3 kernel, for example, yields a reduction in speed by a factor of 3, equivalent to an overhead of 1/3. While this reduction can be avoided with a number of approaches in order

to perform convolutions with a symmetric stride, thus avoiding any reduction in speed, this in fact is actually not necessary for most applications. Finally, our method is extremely flexible as well as highly reconfigurable without the need to change any hardware. We show that the same hardware that is used for image processing with the convolutional accelerator can also be used as an deep learning optical CNN. We demonstrate the performance of this CNN system in separate experiments. The hardware for the convolutional accelerator thus forms the basis of both the input convolutional accelerator processor as well as the subsequent fully connected CNN neuron layer (see below). The network achieves matrix multiplications by sampling each slot of the output waveform, because the vector dot product is equal to the special convolution of two input vectors W and X having the same lengths.

A detailed photonic convolution accelerator that operates in two distinct modes is shown in Figure 3. The left panel illustrates the system when it is operating in convolutional accelerator mode, suitable for both stand-alone large scale convolutional image processing as well as serving as the front end convolutional input of the CNN. The right side of Figure 3 depicts the architecture of the network when it is performing matrix operations for the fully connected optical CNN layer. Since the experimentally demonstrated networks that we use are too complex to be clearly presented, in Figure 3 we show simplified versions of the weights and input data in order to illustrate the system operation principle. The lengths of X and W that we use in Figure 3 are L = 13 and R = 4 for convolutions, and R = L= 4 for matrix operations of the fully connected layer, respectively.

The photonic convolution accelerator schematic is shown on the left side of Figure 3. First, the input vector data of length L and the weight vector of length R are multiplexed in the time and wavelength dimensions, respectively. The input data vector is contained in the temporal symbol intensities of a consecutive electrical waveform X[n] (n represents the temporal positions of the discrete symbols, n ∈ [1, L+R−1]), where X[n] is the accelerator electrical input. The weight vector of the kernel is then imposed on the optical power of the shaped microcomb lines, represented by W[R−i+1], for the i^{th} wavelength channel (i ∈ [1, R], with i increasing with wavelength). The input electrical waveform X[n] is first multicast across all of the the shaped microcomb lines using an electro-optic modulator. Hence, the weighted replica of the i^{th} wavelength channel is given by W[R−i+1]· X[n]. Next, the optical signals of all channels, or wavelengths, are sequentially shifted in time by using an optical time-of-flight buffer. This produces a wavelength-dependent delay which effectively introduces a dispersion or time-step delay between adjacent wavelengths of τ (the difference in delay between wavelengths), which is designed to be equal to the duration of each symbol (the reciprocal of the Baud rate) of X[n]. Hence, the shifted replica is given by W[R−i+1]· X[n-i]. Finally, the replicas for all of the wavelengths are summed together by the photo-detector, and the result is given by

$$\mathbf{Y}[n] = \sum_{i=1}^{R} \mathbf{W}[R-i+1] \cdot \mathbf{X}[n-i] = (\mathbf{W} * \mathbf{X})[n] \tag{1}$$

where each calculated symbol Y[n] within the range of [R+1, L+1] represents the dot product between the kernel W and a region of X given by the sliding receptive field, [n−R : n−1] or [n−R, n−R+1, n−R+2, …, n−1]). By reading the different time slots of the output waveform, a convolution is obtained between the input data and the weight vector. This generates the feature maps, or output matrix convolutions, of the input image. Note that any high order dispersion of the dispersive delay will tend to degrade performance. In our experiments we paid particular attention to this to ensure that it was not an issue.

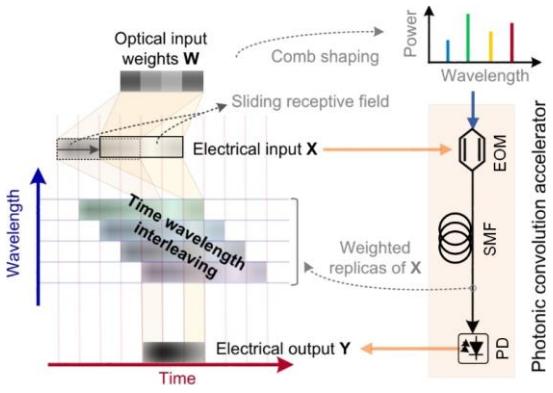

Figure 1 | Principle of operation of the photonic convolution accelerator. EOM: electro-optical Mach-Zehnder modulator. SMF: standard single mode fibre for telecommunications. PD: photodetector.

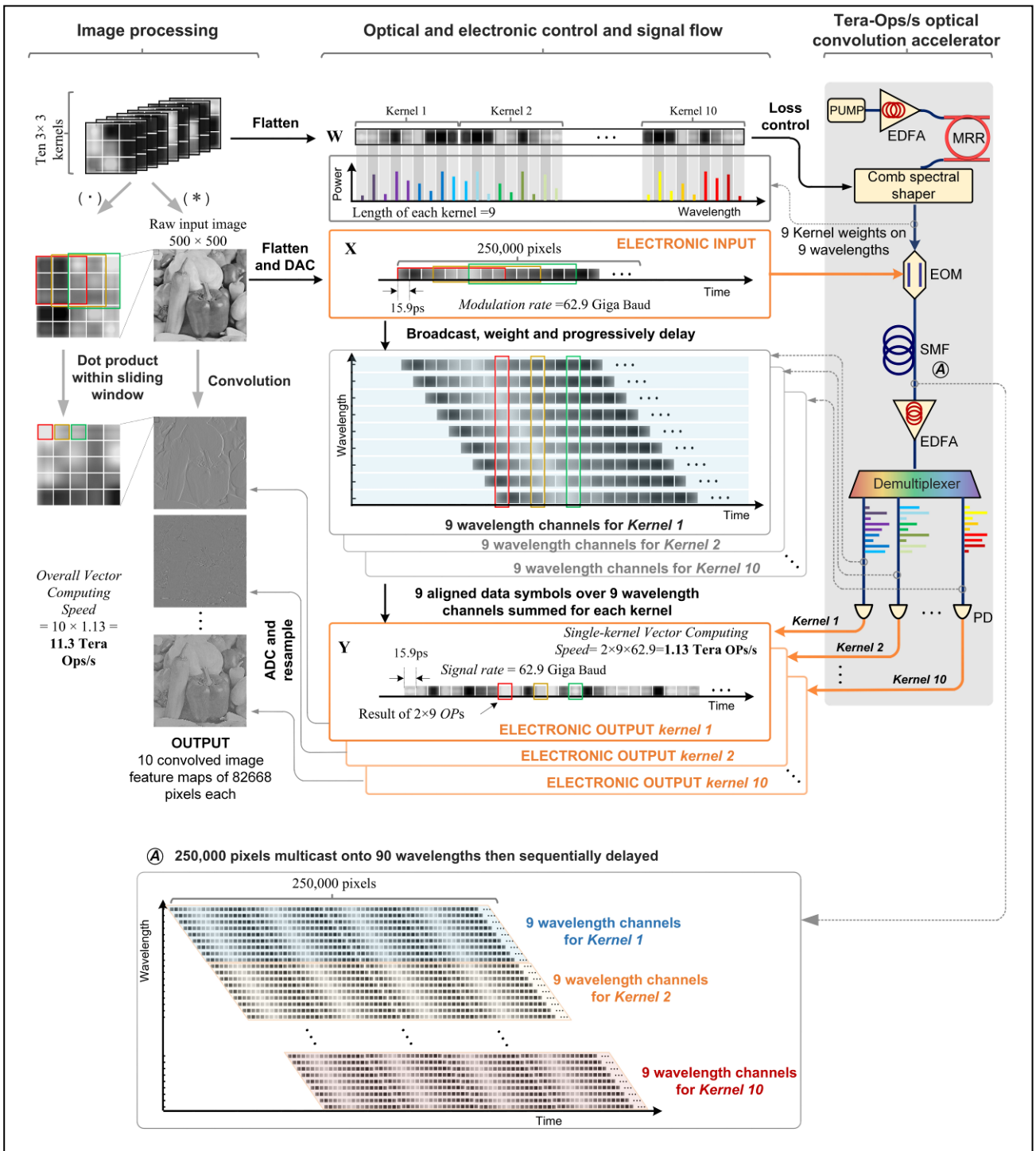

Figure 2 | Image processing, including the experimental setup of the convolutional accelerator (right panel), the signal flow and the optical and electronic control (middle panel), and the processing flow of the raw input image (left panel). PC: polarization controller. MRR: micro-ring resonator. EOM: electro-optical Mach-Zehnder modulator. SMF: standard single mode fibre.

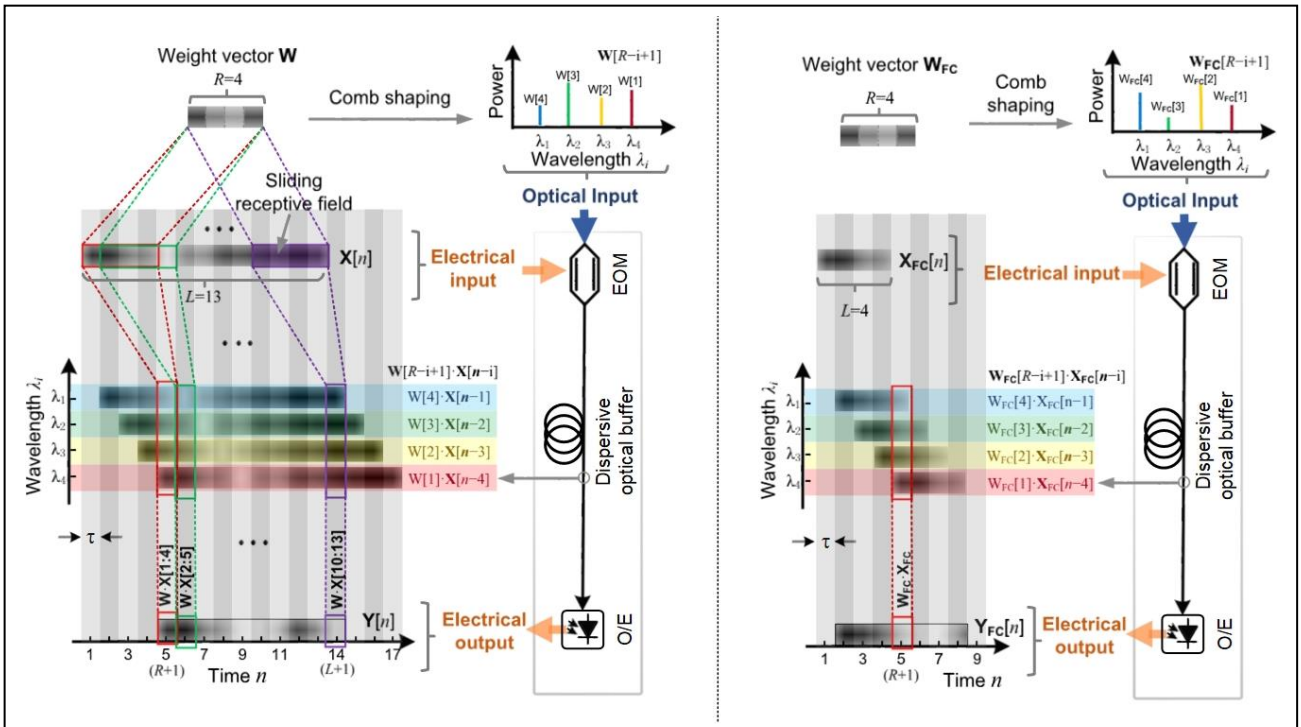

Figure 3 | Photonic convolution accelerator operating in two different modes. Left: convolution accelerator front end, shown for R = 4 and L = 13; Right: matrix operation mode with R = L = 4. Electro-optical modulator (EOM), optical buffer that has progressive wavelength-sensitive delay, and an optical-to-electrical conversion module (O/E).

The convolutional accelerator can also perform matrix multiplications, as shown on the right side of Figure 3. These operations can be viewed as a special case for convolutional operations where the two input vectors, given by the flattened and pooled feature maps, together with the flattened synaptic weights of the fully connected layer, are equal in length (R=L). Figure 3 depicts a case where R=L=4. For this case, we assume that the input data vector, given by XFC[n], and the weight vector given by WFC[R−i+1], both have the same length given by R (i ∈ [1, R], n ∈ [1, R]). Therefore, from Eq. 1, the output waveform following photodetection is given by

$$\mathbf{Y}_{FC}[n] = \sum_{i=1}^{R} \mathbf{W}_{FC}[R-i+1] \cdot \mathbf{X}_{FC}[n-i] \tag{2}$$

By sampling at the time slot denoted by n=R+1, the matrix multiplication of the two input vectors is then given by

$$\mathbf{Y}_{FC}[R+1] = \sum_{i=1}^{R} \mathbf{W}_{FC}[R-i+1] \cdot \mathbf{X}_{FC}[R+1-i] = \sum_{i=1}^{R} \mathbf{W}_{FC}[i] \cdot \mathbf{X}_{FC}[i] \tag{3}$$

Given that convolutional accelerators basically operate on vectors, for image processing applications the data input is in the form of a matrix and therefore it must be flattened into a vector. Here, we use a common technique to do this where the input matrix is first horizontally sliced into sub-matrices, each having a height equal to the convolutional kernel. The sub-matrices are then flattened into vectors and concatenated head-to-tail in order to produce the resulting vector. This image processing flattening process acting with the CNN [14] results in the receptive field sliding with a horizontal stride of 1 and vertical stride of the height of the convolutional kernel. We note that a small stride (such as a horizontal stride of 1) guarantees that all raw data features are extracted, whereas a large vertical stride of 3 or 5 will reduce the overlap between the sliding convolution windows and so will subsample the convolved feature maps. Hence this process will also partly serve as a pooling function. The AlexNet [144 – 146] used a stride of 4.

Further, we mention that while homogeneous strides are often used for digital CNNs, inhomogeneous convolutional strides having unequal vertical and horizontal strides, as we use here, are very commonly used and do not limit the convolution accelerator performance. For our work, this was verified by the high recognition accuracy we with the CNN for the full 10 digit set. Furthermore, if the homogeneous convolutions can be accomplished by replicating the weight-and-delay paths, with each one containing the modulator, fibre spool, de-multiplexer and photo-detector, this can be completely avoided.

3. EXPERIMENT

3.1 Soliton crystal optical frequency micro-combs

Optical frequency combs, with arrays of equally spaced discrete frequencies lines, are extraordinarily useful for many applications - particularly optical frequency metrology [29]. Micro-combs provide the full capability of mainframe optical frequency combs, but in a compact and integrated form having an extremely small footprint [29-35]. Micro-combs have underpinned a great number of ground-breaking achievements in many areas, including optical frequency high-resolution synthesizers [33], ultrahigh bit rate telecommunications [34, 35], high dimensional and complex quantum state generation [36 - 44], high performance microwave and RF signal generation, detection, and processing [68 - 88], and much more. Figure 4 shows an illustration of the micro-comb chip together with output spectra with the accompanying optical pumping curves. Our work is based on a type of micro-comb that has been called "soliton crystals" because they exhibit a profile that resembles a crystal structure with tightly packed self-localized pulses in the angular dimension in the ring [35, 48, 49]. They naturally form in micro-cavities that have particular dual mode crossings, and most importantly do not require any complicated pumping dynamics, nor do they require any external stabilization methods. They are governed by the Lugiato-Lefever equations [29, 47] and exhibit unique optical spectra (Fig. 4f) profiles typically termed "palm shaped", resulting from the spectral interference between the tightly packed solitons that circulate around the micro-ring. Soliton crystals obey highly reliable and repeatable deterministic generation that results from interference induced by the mode crossing on the background wave and the very high optical power in the cavity (Fig. 4c). This allows a very simple and reliable initiation even with just simple and slow adiabatic pump wavelength tuning. [35] In fact, soliton crystals can even be generated by using simple manual tuning of the pump laser wavelength. Figure 4d shows the intracavity optical power inside the micro-ring during the pump sweeping process. The reason that it is possible to slowly or adiabatically sweep the pump is because the optical power inside the cavity is over thirty times higher than that of single dissipative Kerr solitons (DKS), to the point where it is very close to the power of the background spatio-temporal chaos states [29, 35] from which they emerge. Therefore, soliton crystals have negligible thermal detuning that, for the case of DKS states, arises from the 'soliton step' at threshold and which leads to instability. This makes resonant pumping of DKS states extremely difficult, requiring dynamic pump wavelength and power "kicking" often with reverse tuning. It is this combination of high conversion efficiency together with the ease of generation that results in soliton crystals being highly useful. Our coherent optical soliton crystal microcomb (Figure 4) was a result of optical parametric gain followed by oscillation in a single integrated MRR (Fig. 4a, 4b). The device was fabricated in CMOS-compatible Hydex [23, 24, 35], and achieved a Q factor greater than 1.5 million, with a radius of 592 μm, and a very low free spectral range (FSR) of ~ 48.9 GHz. The pump laser power was amplified by an erbium doped optical fiber amplifier (Pritel PMFA-37) to produce soliton crystal microcombs with more than 90 channels across the telecommunications C-band (1540-1570 nm). It resulted in very low-noise frequency microcomb lines produced by a device having an extremely small size of less than one square millimeter, with a power consumption of less than 100 milliwatts using the technique in [35].

3.2 Matrix Convolution Accelerator

The experimental arrangement for the convolutional accelerator operating in full matrix mode in order to operate on classic 500×500 face image is shown in Figure 2. The system generates 10 parallel convolutions simultaneously with 3×3 kernels in order to produce the distinct image processing functions. The weight matrices for all ten kernels were flattened into a composite kernel vector represented by W that embodies all of the 90 weights (10 kernels with 3x3=9 weights each). These were all then encoded onto the optical power of all ninety microcomb lines through the use of an optical spectral shaper (Waveshaper). Each kernel occupied its own frequency band consisting of nine wavelengths. The wavelength channels were generated by a coherent soliton crystal microcomb source that operated via optical parametric oscillation within a single micro-ring resonator (MRR) (Fig. 4b) with a radius of 592 μm, a FSR spacing ~ 48.9 GHz and with an optical bandwidth of ~ 36 nm for 90 wavelengths in the C-band (1540-1570 nm) [35].

The experimental results of the image processing are shown in Figure 5, where Fig. 5a shows the kernel weights and the optical spectrum of the shaped microcomb, while at the same time the electrical input waveform of the image (grey lines are theoretical while the blue lines are the experimental waveforms) are in Figure 5b. Figure 5c depicts the convolved results of the fourth kernel that consists of a top image processing Sobel function (grey lines = theory, red = experimental). Finally, Figure 5d displays the matrices of the corresponding recovered images and weight kernels.

The raw input face image that is 500×500 pixels in size was first electronically flattened into a vector X and then encoded as the intensities of the 250,000 temporal symbols, each having a resolution of 8 bits/symbol (limited by the electronic arbitrary waveform generator (AWG)), in order to form the electrical input waveform through the use of a high-speed electrical digital-to-analog converter, at a data rate of 62.9 Giga Baud (time-slot τ =15.9 ps) (Fig. 5b). The duration of each waveform was 3.975μs for each image, yielding a processing rate for all ten kernels of > 1/3.975μs, equivalent to 0.25 million ultra-large-scale images a second.

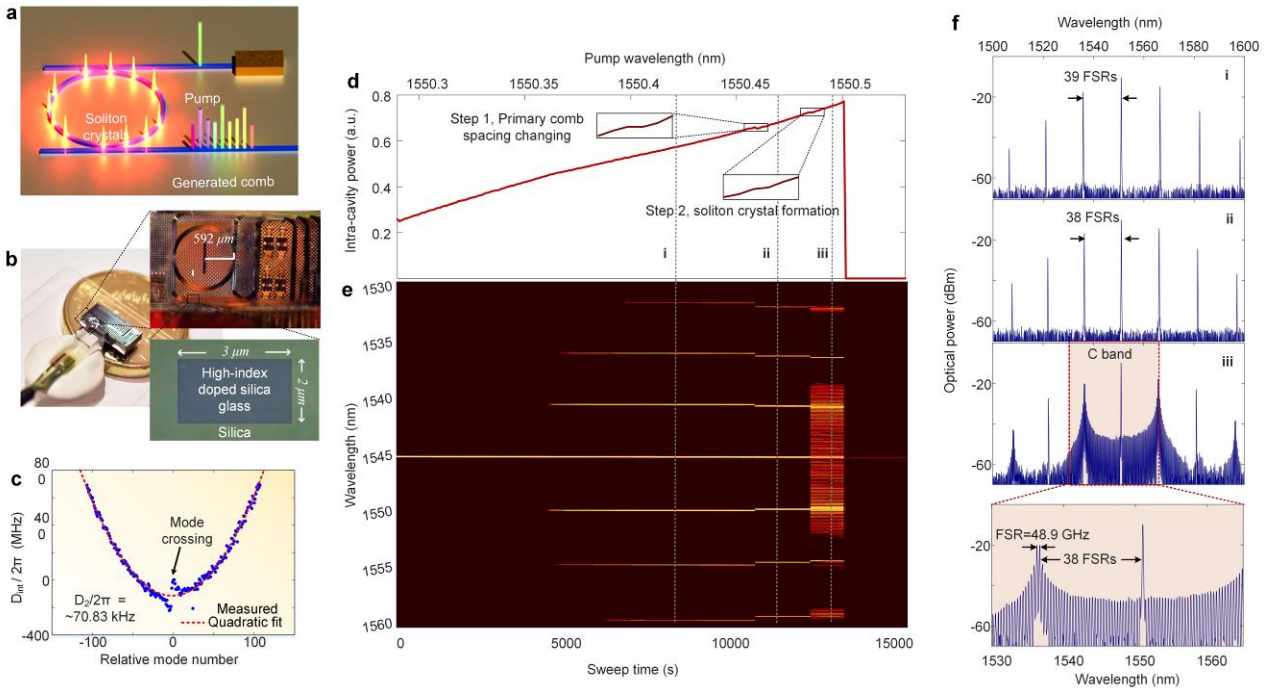

Figure 4 | a. Soliton crystal microcomb, generated by pumping an on-chip high-Q nonlinear micro-ring resonator with a CW laser. **b.** Image of the MRR and a SEM image of the MRR cross section. **c.** Measured dispersion of the MRR showing the mode crossing at ~1552 nm. **d.** Measured soliton crystal step of the intra-cavity power, and **e.** optical spectrum of the microcomb when sweeping the pump wavelength. **f.** Optical spectrum of the generated coherent microcomb at different pump detunings at fixed power.

The input waveform (X) was then multi-cast onto all ninety shaped microcomb lines using an electro-optic modulator. This yielded replicas that were each weighted by the kernel vector W. Following this, the waveform was propagated across about 2.2 km of standard single mode optical fibre that had a standard dispersion equal to 17ps/nm/km. This length of fibre was chosen carefully to yield a sequential time shift in the weighted replicas with a progressive delay step of 15.9 ps between the adjacent wavelengths. This precisely matches the length of each input data symbol τ, and so results in wavelength to time interleaving for all of the 10 kernels.

After this, the ninety wavelengths were de-multiplexed into ten sub-bands each consisting of 9 wavelengths, with every sub-band equating to each kernel, and with each separately detected by ten high speed photodetectors. Detecting each wavelength acts to sum the temporally aligned symbols of each replica, and the output electrical signal for each kernel is shown in Fig. 5c (for kernel 4). All 10 electrical waveforms for the kernels were then converted into digital waveforms using ADCs and then resampled so that each waveform time slot contained the vector dot product between a kernel convolutional matrix and the input image, all moving with a sliding receptive field, or window. The output signals thus contained the ten convolutional matrix outputs (feature maps) representing the hierarchical extracted features of the input image waveform (Figure 5d).

The convolutional accelerator was based on the combination of spatial wavelength and temporal multiplexing, with the convolution window essentially sliding across X, the input data signal, at 62.9 Giga Symbols/s, the modulation baud-rate. Each output symbol is generated by nine MAC (multiply-and-accumulate) operations, which is the size of each of the kernels. Therefore, the throughput, or vector computing speed for each kernel is 2×9×62.9 = 1.13 TOPS, and since all 10 kernels were generated simultaneously in parallel, the net computing speed of the system is given by 1.13×10 = 11.3 TOPS, or 11.321×8=90.568 Tb/s. Note that the speed in terms of Tb/s is slightly reduced by the optical signal to noise ratio (OSNR), which acts to reduce the effective number of bits (ENOBs), but in our case this reduction was small. Note that this speed is over 500 times the fastest ONNs that had been reported when this work was first published, and only one system published since has been faster [16].

For the experiments on matrix image processing reported here, the convolution window had a sliding vertical stride of 3 (because of the kernels being 3×3), and so the net matrix computing speed was reduced by 3 to 11.3/3=3.8 TOPs. The use of homogeneous strides with unity horizontal and vertical strides, that would result in full vector speed operation, can easily be accomplished simply by replicating the system in parallel with weight-and-delay paths [14]. As mentioned, however, we did not find this necessary in order to achieve our successful performance. For the experiments performed here, the input data had a length of 250,000 pixels, but in principle this is unlimited in size – the only consideration is the practical limitation imposed by the capacity of the external electronics.

To implement the required weights for the kernels, the power of the soliton crystal microcomb lines were modulated through the use of 2 commercial spectral shapers (Finisar WaveShaper 4000S) that are based on liquid crystal on silicon (LCOS) technology. The first waveshaper flattened the microcomb lines while the second was positioned immediately before the photo-detectors, and accomplished the precise shaping of the come lines that was required by the kernel weights. We used a feedback loop to increase the comb shaping accuracy with an error signal being obtained by measuring the system impulse response for an input pulse (Gaussian shaped) and comparison with the theoretical weights.

Both the theoretical and experimental facial image processing results produced by the matrix convolutional accelerator for all 10 kernels, is shown in Figure 6, for the 500×500 pixel (250,000) facial image, featuring both the recorded waveforms and the recovered images. The time dependent waveform at the operation baud rate was generated from the electrical input data through the use of an electronic arbitrary waveform generator (AWG) (Keysight M8195A). The electrical waveform was then converted to optical signals by multicasting it onto all wavelength channels simultaneously using an electro-optic intensity modulator (iXblue) having a bandwidth of 40GHz. We used sampling points at a speed of 62.9 Giga samples/s for the image processing to form the input symbols. All of the optical signals, propagating down the same single fibre, were then transmitted across 2.2 km of standard single mode fiber with a dispersion of 17ps/nm/km in the telecom band , that was chosen to generate a relative temporal delay between adjacent wavelengths 15.9 ps/channel, in order to exactly match the baud rate of the input data.

Since there are no conventions in this very young field for quantifying or classifying the processing computing speed or performance of ONNs, here we clearly outline the definitions for performance that we use to quantify the system performance. We adapt a method that is commonly used in the electronics world to evaluate the performance of micro-processors. Here, the convolution accelerator processing power, which is closely related to the operational bandwidth, is termed the data throughput, given by the number of operations performed within a given time period. Since in our system the weight vectors and input data follow different paths and are interleaved in temporal, wavelength, and spatial dimensions, we use the time sequence at the electrical output port to calculate the throughput in a clear manner.

At the electrical output port, the output waveform has a total of L+R−1 symbols (where L and R are the lengths of the data input vector and kernel vector weights, respectively), amongst which L−R+1 symbols consist of the results of the convolution. Each output symbol is the result of R MAC operations or 2R OPS, where the duration of the symbols τ is given by the waveform input symbol duration. Therefore, since L is typically much greater than R for practical neural networks, the term (L−R+1)/(L+R−1) does not affect the throughput, or vector computing speed, which is given by (in OPS)

$$\frac{2R}{\tau}\cdot\frac{L-R+1}{L+R-1}\approx\frac{2R}{\tau} \tag{4}$$

Hence, the computing speed of the vector convolutional accelerator that we report here is given by 2×9×62.9×10 = 11.321 Tera-OPS, or TOPS, for 10 parallel convolutional kernels.

Many applications, such as speech or audio recognition, operate directly on vectors and in that case the computing speed of the system is given directly by the vector speed of 2R/ τ. On the other hand, when processing matrix data as is typical for images, one has to include the overhead that reduces the net computing speed as a result of the flattening process used to convert the matrix to a vector. This is directly related to the size of the convolutional kernel, with 3-by-3 kernels, for example, generating an overhead of 1/3 so that the effective computing speed is correspondingly reduced to ~ 1/3 * 2R/τ. In our case this still resulted in a matrix operation speed that was in the multiple TOP regime, a result of the high degree of leveraging and parallelism yielded by interleaving the spatial-temporal-wavelength dimensions.

The output waveform for each kernel of the convolutional accelerator (length = L−R+1=250,000−9+1=249,992) contains 166×498=82,668 useful symbols. These are then temporally sampled in order to obtain the feature maps, with the balance of the symbols being discarded. Hence, the net experimental speed of the matrix convolution process is slower than the corresponding vector speed by the overhead factor of 3, with a resulting net speed of 11.321 × 82,668/249,991 = 11.321 × 33.07% = 3.7437 TOPS.

For the deep learning CNN, the front end convolutional accelerator layer had a computing vector speed of 2×25×11.9×3 = 1.785 TOPS while the convolution matrix speed, where we use 5x5 kernels, was 1.785×6×26/(900−25+1) = 317.9 Giga-OPS. For the deep CNN fully connected layer, Eq. (4) shows that each neuron's waveform output has a length of 2R−1, with the relevant (ie., useful) symbol being the one located at R+1, also arising from 2R operations. Hence, the fully connected layer speed is 2R / (τ*(2R−1)) for each neuron, and so for R = 72 for the experiment, and with 10 neurons being calculated in parallel, the net matrix multiplication speed then becomes 2R / (τ*(2R−1)) × 10 = 2×72 / (84ps* (2×72−1)) = 119.83 Giga-OPS for our experiments.

One further consideration is that the digital system intensity bit resolution in terms of effective number of bits (ENOBs) for analog ONNs is affected by the signal-to-noise ratio (SNR). For an ENOB of 8-bits, the system SNR must be greater

than 20·log10(28) = 48 dB. Our system had an SNR greater than this and so our speed in Tb/s was very close to 8 x the speed in Ops, and not reduced noticeably by our OSNR.

3.3 Deep Learning Optical Convolutional Neural Network

Our convolutional accelerator is scalable as well as completely and dynamically reconfigurable without any changes to the hardware. Hence, we used the exact same hardware to achieve the stand alone accelerator as well as both components of the CNN – namely the front-end convolution accelerator followed by the fully connected layer, which both together comprised the deep optical CNN. We used the system (optical CNN) to perform recognition of the full ten digit (0-9) handwritten images, a task that is rarely demonstrated with optical networks. Figure 7 shows the full deep (multiple) level optical CNN architecture. The feature maps are the convolutional matrix accelerator front end generates the feature map outputs that are then input into the subsequent fully connected layers that embody the neural network of the system.

The optical CNN architecture (Figure 8) includes convolution, pooling, and fully connected layers. Figure 9 shows The detailed experimental schematic of the optical CNN (Figure 9) shows the input front end convolutional accelerator on the left side and the fully connected layer on the right - both together forming a deep learning (more than one layer) optical CNN. The optical wavelengths were generated by the soliton crystal microcomb, that were used by both the convolution accelerator and fully connected layers. A digital electronic signal processing (DSP) module was used for the temporal sampling, with the pooling function being performed by an external computer, although this latter function can be readily done all-optically.

The convolutional layer generates the heaviest computing workload of the entire network, typically occupying anywhere from 55% to 90% of the total system processing power. The digit images consist of 30×30 matrices each, containing grey-scale quantities that are digitized with a resolution of 8-bits. These were then flattened into vectors and then converted into a temporal signal at a speed of 11.9 Giga Baud (time-slot τ =84 ps) – about 5 times slower than the stand alone convolutional accelerator discussed previously. We employed three kernels, each of size 5×5, and so the system needed 75 wavelengths from the microcomb. This resulted in a vertical stride of 5. The required temporal dispersive delay was generated by about 13 km of SMF in order to be commensurate with the input data baud-rate. The microcomb wavelengths were separated, or de-multiplexed, into the 3 kernels, each of which were then photodetected separately with high speed photodetectors, after which the electronic waveforms were sampled and finally scaled with a nonlinear transfer function using digital electronics to generate the output of the convolutional front-end which consisted of the extracted hierarchical feature maps of the input images. The feature maps were then pooled electronically and flattened into a vector (Eq. 2,3) X_{FC} (72×1= 6×4×3) per image that was then sent to the fully connected layer.

The fully connected layer featured 10 neurons, one each for each of 10 handwritten digits (0 to 9). The input consisted of synapses with the synaptic weights represented by a 72×10 weight matrix $\mathbf{W}_{FC}^{(l)}$ (ie., 10 column vectors each of size 72×1) for the l^{th} neuron ($l \in [1, 10]$) – with the number of microcomb wavelengths (72) equal to the length of the flattened feature map vector $\mathbf{X}_{FC}$. The l^{th} port had a shaped optical spectrum with the optical power distribution being determined by the weight vector $\mathbf{W}_{FC}^{(l)}$, which therefore served as the effective optical input to the l^{th} neuron. After multicasting onto all 72 wavelengths simultaneously with a single optical modulator, and then propagating through the single mode fiber which generated the sequential delays between wavelengths, the optical signals were then weighted and then spatially and spectrally demultiplexed into 10 spatial output ports using a single Waveshaper — each corresponding to one neuron. Since this part of the network operated via linear processing, the kernel wavelength weighting can be performed at any stage of the network - either before EO modulation or later just in front of the photodetectors. The advantage of applying the weights at the end just before detection is that both the demultiplexing and weighting functions can be performed using only one Waveshaper rather than two. Finally, the different outputs of the nodes and neurons were extracted through sampling the 73rd symbol of the convolved results. The final output of the optical CNN was contained in the intensities of the output neurons, with the neuron having the highest intensity yielding the predicted category of the tested image. The external, or peripheral systems used in our experiments included signal sampling, the nonlinear transfer function and the pooling functions. Here, these functions were performed with electronic digital signal processing hardware. However, in principle most, if not all, of these functions could be achieved purely in the optical domain with the VCA. The supervised network training, which only needs to be performed once for a given network task, was achieved beforehand using offline digital electronics.

Using the deep optical CNN, we experimentally tested 500 images, each having a resolution of 8-bits, with a pixel matrix size of 30 × 30 pixels (900 pixels) for each of the handwritten digit dataset. Figure 10 shows the resulting confusion matrix that indicates that our system has a prediction accuracy of 88%, versus a theoretical accuracy of 90% calculated by computer. The computing speed of the CA front end of the system was 2×75×11.9 =1.785 TOPS, or 14.3 Tb/s. As mentioned, to perform the convolutions of the image matrices we used 5×5 kernels, yielding a matrix flattening overhead of 5, and net speed of 1.785/5= 357 Giga OPS. The computing speed of the fully connected layer was 119.8 Giga-OPS. The waveform duration was 30×30×84ps=75.6ns for each image, and so the convolutional layer processed images at a rate of 1/75.6ns = 13.2 million digit handwritten images a second.

We note that full 10 digit (0 - 9) handwritten digit recognition, although commonly performed by electronic digital networks, is very rare for optical ONNs. It requires a very large number of parallel physical paths for fully connected networks. For example, a hidden layer having 10 neurons would need 9000 physical paths, which represents a significant challenge for state-of-the-art nanofabrication technologies. Our CNN is the first integrable and reconfigurable ONN that able to not only perform complex and high level operations, including the full 10 handwritten digit recognition, but also at Tera-OP per second (TOP) processing speeds. The CNN convolutional layer used 5 sample points at 59.421642 Giga Samples/s to generate each single symbol of the input waveform, which matched the progressive delay time (84 ps) generated by the 13km long dispersive fibre. The generated electronic waveforms for 500 testing images [14] formed the electrical signal input of both the convolutional and fully connected layers.

In both the 500×500 image processing experiment and the convolutional layer of the CNN, for the convolutional accelerator component the second Waveshaper both de-multiplexed and shaped the wavelength channels into the different separate spatial ports according to the configuration of the convolutional kernels. For the fully connected layer, the second Waveshaper simultaneously performed both shaping and power splitting (not de-multiplexing) for the 10 output neurons. The de-multiplexed or power-split spatial ports were sequentially detected and measured. However, these two functions could readily be achieved simultaneously in parallel with a commercial 20-port optical spectral shaper (WaveShaper 16000S, Finisar) using multiple photodetectors. We achieved the negative channel weights by one of two methods. For processing the 500×500 image using the convolutional layer of the CNN, the wavelength channels for each kernel were divided into two separate spatial outputs by the WaveShaper depending on the sign of the kernel weight, and then detected using a balanced photodetector (Finisar XPDV2020). On the other hand, the weights for the fully connected layer were encoded in the input electrical waveform symbols by the electronic digital processing stage. Both of these approaches towards implementing negative weights were successful. Finally, the electrical output waveform was sampled and digitized by a high-speed oscilloscope (Keysight DSOZ504A, 80 Giga Symbols/s) to derive the final convolved output. The extracted outputs of the convolution accelerator part of the CNN were additionally processed digitally, including rescaling to factor out the photonic link loss using a reference bit, then mapped onto a determined range with a nonlinear tanh function. The pooling layer was also performed digitally, using an algorithm introduced in the network model. The residual small disagreement between theory and experiment, for both the recognition and convolution functions, arose mainly from the degradation of the input waveform arising from the electrical arbitrary waveform generator's performance limitations. Improving this would produce greater accuracy and result in better agreement with theory.

3.4 Network training and digital processing

For the deep learning optical CNN experiments we used the MNIST (Modified National Institute of Standards and Technology) datasets of handwritten digital images [144] that comprises 60000 images for the training set and 10000 images as the test set. Figure 7 shows the structure of the CNN. In our case designing the network was significantly helped by the fact that the number of synapses and neurons of the network could be dynamically reconfigured without the need to modify the hardware. The 28×28pixel images of the input data was first padded with zeros into 30×30 images and then sliced into 5×180 matrices which were then convolved with the 5×5 kernels. The slicing operation made the receptive field slide horizontally with a stride of 1 across the rows and a vertical stride of 5 over the columns of the 30×30 input data (= 900 input nodes). Following this, the 6×26×3 feature maps were then pooled (using an average pooling method) to create matrices with a smaller dimension of 6×4×3. Finally, the matrix was further flattened into a 72×1 vector that generated the input nodes to the fully connected layer. Finally, the fully connected layer calculated the predictions which were encoded into the output of the 10 neurons. We used a nonlinear tanh function following the convolutional layer, pooling function and fully connected layers. We chose this, rather than other nonlinear functions such as ReLU which are widely used, since it can be experimentally achieved using a saturating electrical amplifier.

The training to determine the pre-trained weights and biases was achieved prior to the experiments, offline using a digital computer. We used the well-known Back Propagation algorithm [145] to adjust the weights. To validate the hyper-parameters of the CNN, we performed a 10-fold cross validation with 60000 samples of the training dataset, separated into 10 subsets and each was then used to test the trained network (6000 samples) with the rest of the 9 subsets (54000 samples). The test sets [14] were processed by both the optical CNN (500 images) and an electronic computer (10000 images) for comparison.

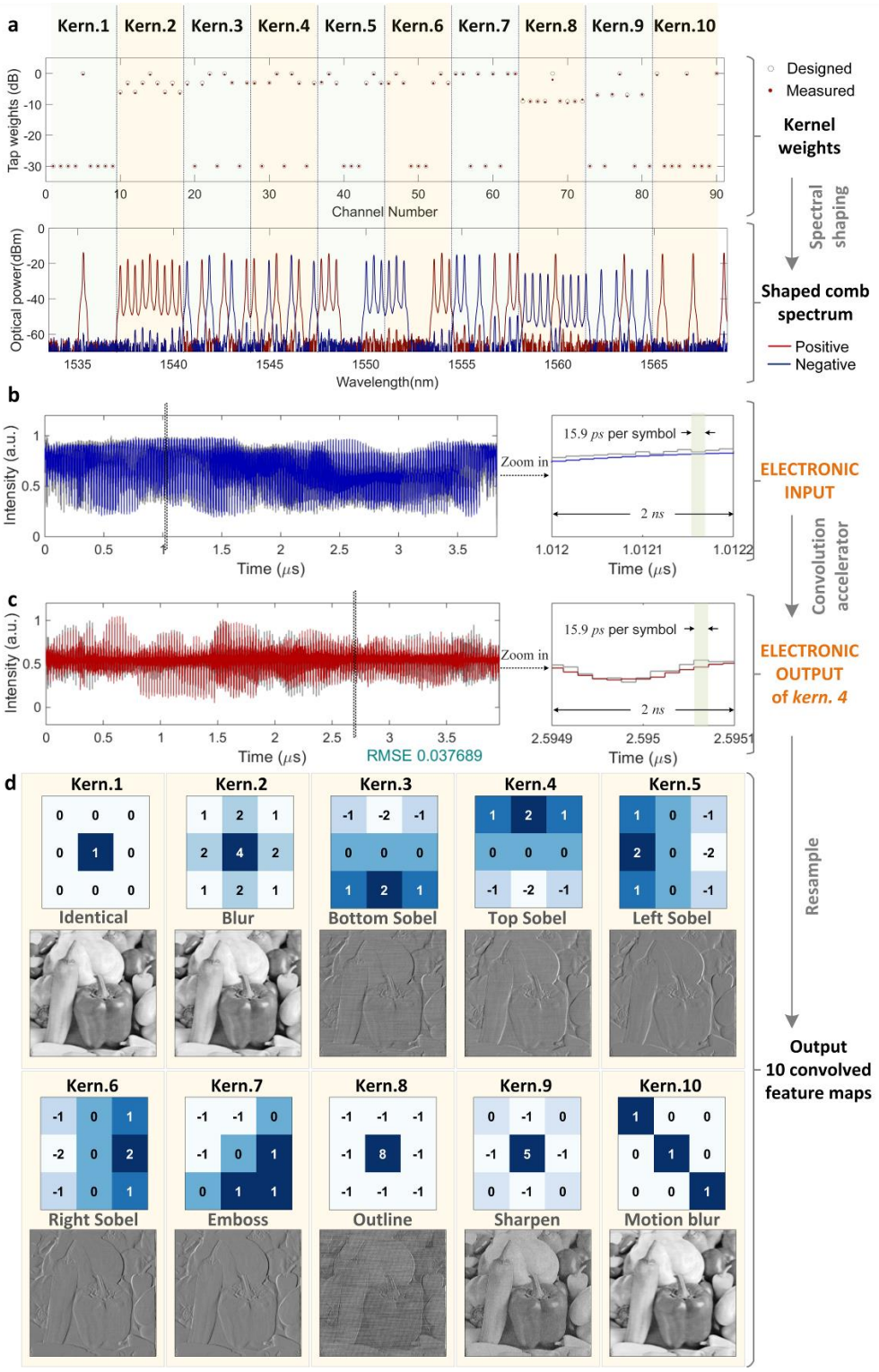

Figure 5 | Experimental results of the image processing. a. The kernel weights and the shaped microcomb's optical spectrum. b. The input electrical waveform of the image (grey lines are theory and blue experimental). c. The convolved results of the 4th kernel that performs top Sobel image processing. d. The weight matrices of the kernels and corresponding recovered images.

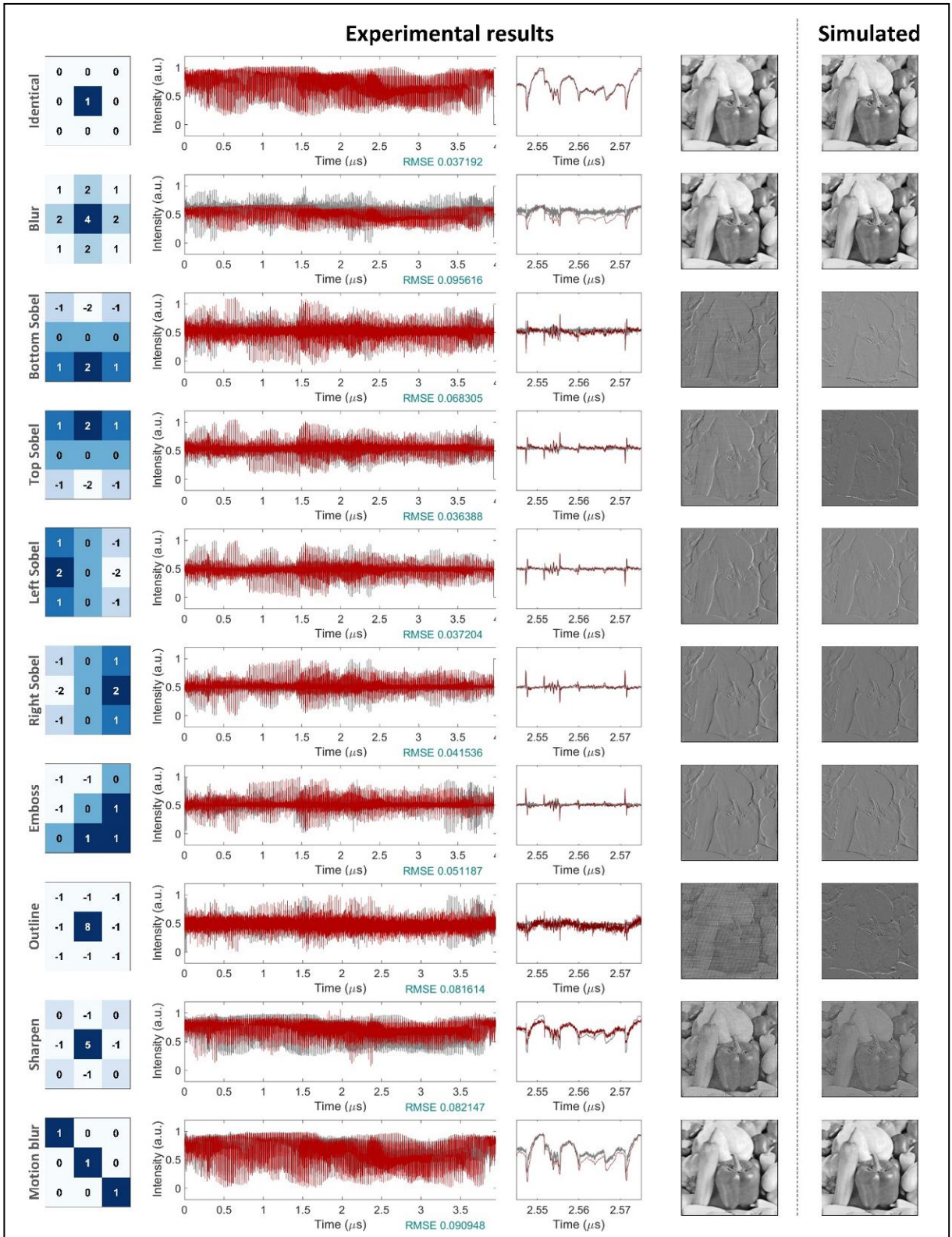

Figure 6. Experimental and simulated large scale facial image processing results performed by the convolutional accelerator with simultaneously processing ten convolutional kernels.

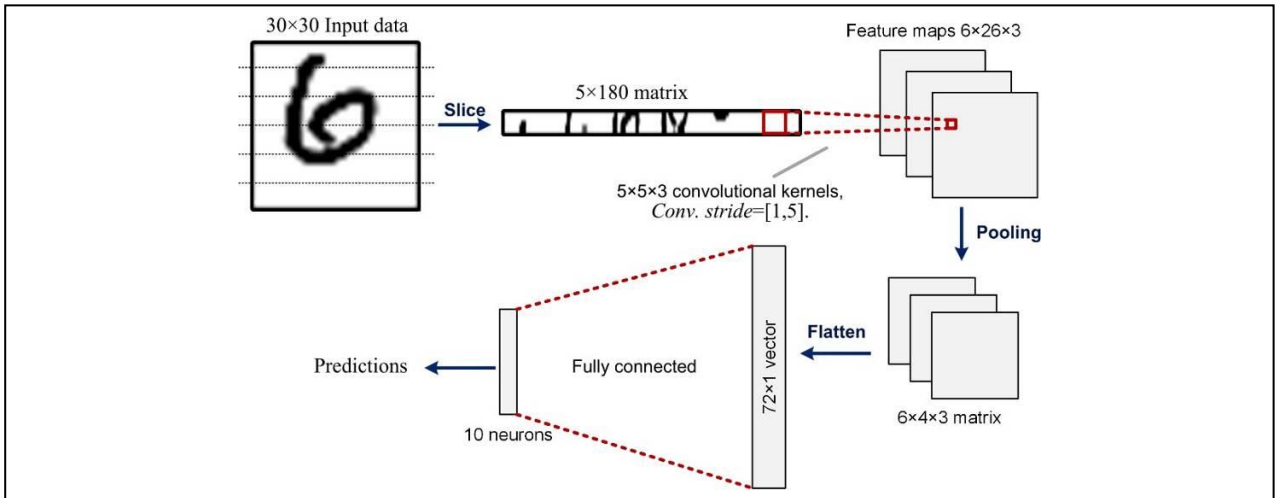

Figure 7. Deep (multiple) level CNN structure. The feature maps are the convolutional matrix outputs while the fully connected layers embody the neural network component.

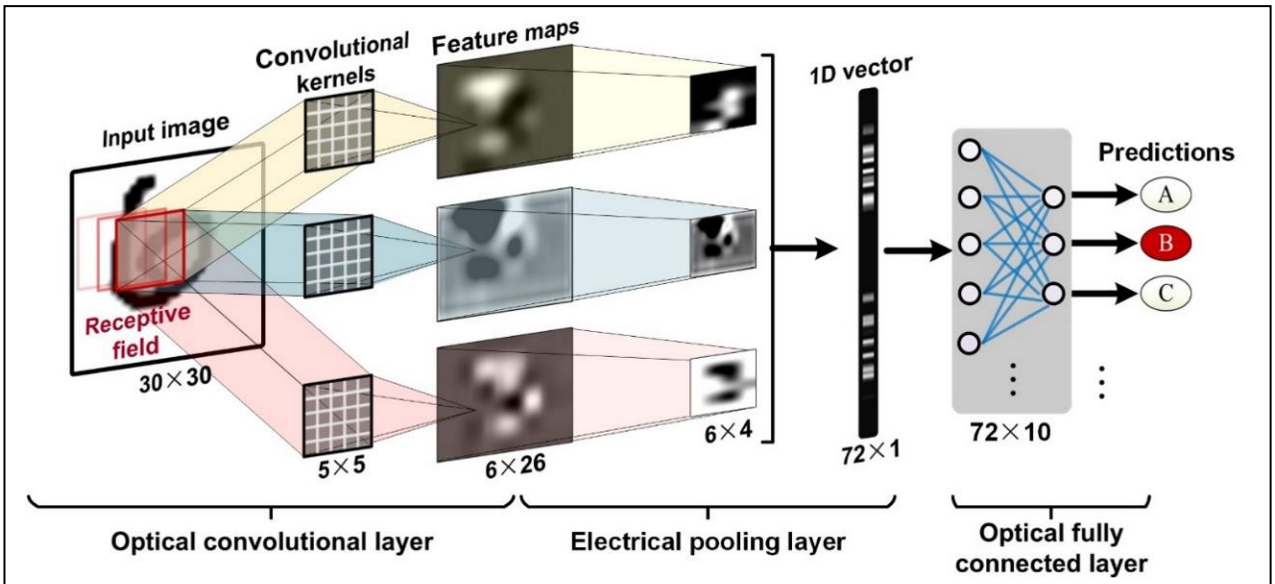

Figure 8. The architecture of the optical CNN, including the convolutional, pooling, and fully connected layers.

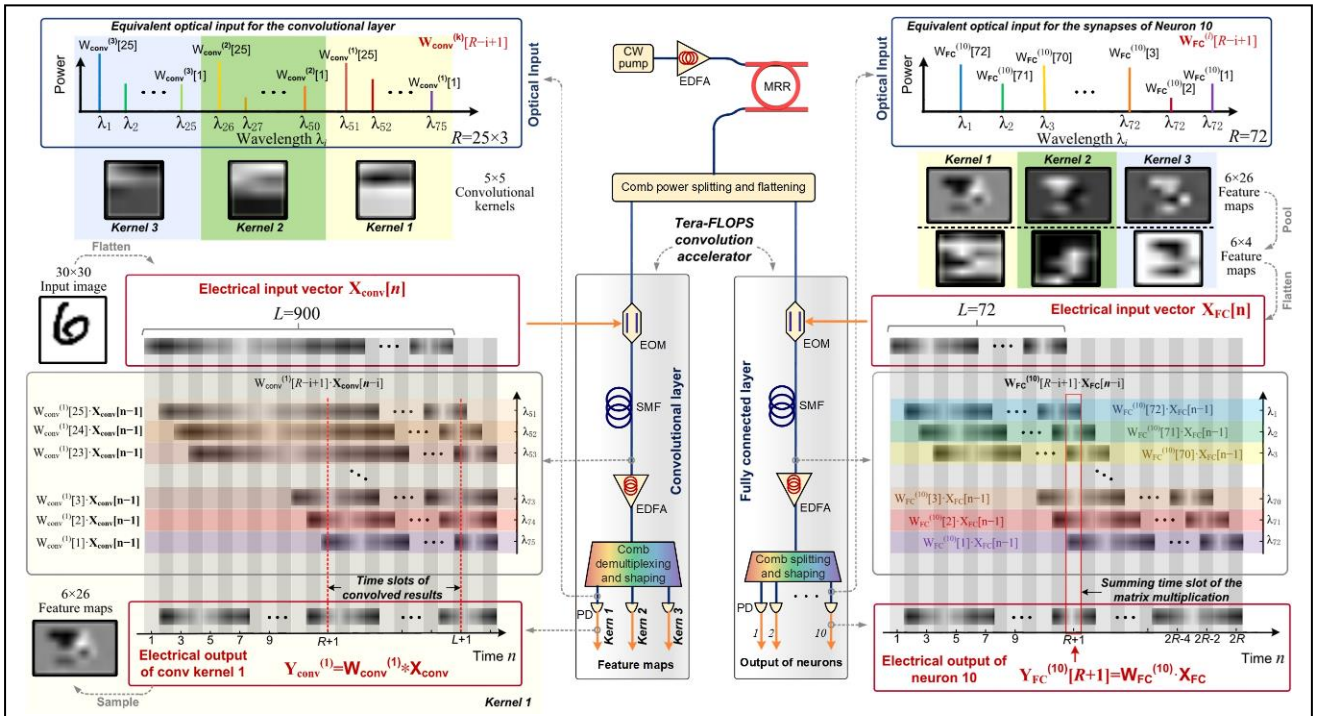

Figure 9. Experimental configuration of the optical CNN. Left side: input convolutional accelerator front end. Right side: fully connected layer, both of which comprise the deep learning optical CNN. The microcomb generates the wavelengths for both the convolution accelerator and fully connected layer. The electronic digital signal processing (DSP) module that performs the sampling and pooling functions is separate to this structure.

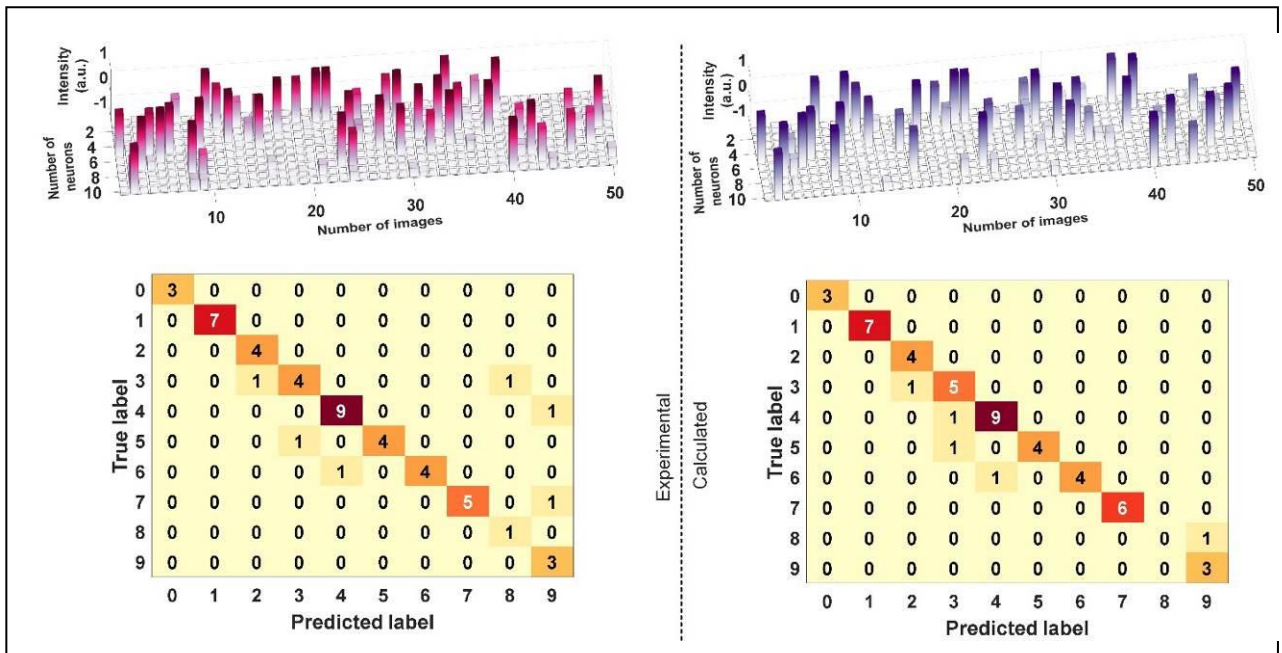

Figure 10 Experimental and theoretical results for image recognition. The upper figures show the sampled intensities of the ten output neurons of the fully connected layer, while the lower figures show the confusion matrices with the darker colours indicating a higher recognition score.

Table 1

Performance comparison of state-of-the-art optical neuromorphic hardware

Parameter / Approach	Input data dimension	Computing speed (OPs/s)	Scalability & reconfigurability*	Integrated components
Free space optics [11]	784	CW §	Level 1	None
Coupler array [8]	4	CW §	Level 2	Weight and sum circuits
Reservoir [9]	371 (up to 1113)	~17.6 G	Level 2	None
Phase change material [12]	15	CW §	Level 2	Weight and sum circuits, nonlinear function
Broadcast and weight [22]	8	~1 G	Level 3	None
Time and wavelength interleaving (this work)	250,000 (theoretically unlimited)	Vector: 11.321 T Matrix: 3.8 T	Level 4	90-wavelength light source

CW§: Indicating the approach used continuous-wave sources as the input data signal, high-speed updating of the input data is not demonstrated to achieve a high computing speed.

3.5 Performance comparison

We list the performance of recent optical neuromorphic processors in Table 1. This section is not intended to be complete but just highlights some key work that addresses the most crucial technical issues for optical processors. The dimension of the input data directly governs the processing task complexity. In real world applications, the dimension of the input data is typically quite large, with a human face image requiring > 60,000 pixels, for example. Therefore, to ultimately make optical computing hardware practical, the input data dimension needs to be at > 20,000 at a minimum. Here, we process of images with 250,000 pixels, which is 224 times higher than previously reported.

The most important metric for computing hardware is arguably the computing speed and this is the main advantage of optical systems. Although to date there has not been a consensus on a definition of optical hardware computing speed, the key parameter is the number of data sets that can be processed within a given time frame - how many images per second can be processed. Therefore, although some systems [8, 11, 12], have a low latency because of the short optical paths, the computing speed is nonetheless very low because of the lack of high-speed data interfaces - in other words, the input and output nodes are not updated at a high rate. Although some methods [9, 28] have high-speed data interfaces, the parallelism of their computing is not high and so the speed is very similar to the input data rate. In our work, [14] we achieve a record speed of 11.321 Tera-OPS, more than 500 x higher than any systems that had been reported when our results were published, by using high-speed data interfaces (62.9 Giga Baud) combined with time-wavelength interleaving,.

Finally, the reconfigurability and scalability of our method greatly enhances the versatility of the optical computing hardware. Approaches that are unable to reconfigure the synapses dynamically [11] (termed “Level 1” in the table) are hardly even trainable. While Level 2 systems [9, 12, 28] support online training, they are only able to perform one specific task since the network structure is fixed and inflexible after the system is fabricated. Level 3 systems [28] can process different tasks although the function of each layer is fixed. This significantly limits the hardware in terms of being able to implement more complex tasks than matrix multiplication. Our work is the first that operates at Level 4 with full dynamic reconfigurability in all aspects of the system performance. With our approach the synaptic weights can be reconfigured simply by software reprogramming of a WaveShaper. Moreover, the number of synapses for each neuron can also be redesigned by re-assigning the wavelength channels using the de-multiplexer. The number of layers can be redesigned by simply varying the number of stacked devices. Finally, the computing function can be changed from convolution to matrix multiplication simply by changing the method of sampling. Finally, the physical level of integration of the system determines the computing density (processing capability per unit area). For systems that are not that amenable to integration [8, 11, 28], the ultimate achievable computing density is very low. While other approaches have achieved a limited level of integration of the weights and summing circuits [8, 12] - probably the most challenging issue — advanced integrated light sources — have yet to be achieved. The performance of the light source directly impacts the capacity of the overall system for both input data scale [8] and number of synaptic connections per neuron [12]. The square millimeter sized microcomb that we use produces a great many precisely spaced wavelengths. This significantly enhances the overall parallelism and thus the computing density. Our approach is a significant step towards the full integration of optical computing hardware.

4. SCALING THE NETWORK

The architecture that we present here based on optical microcombs is highly amenable to scaling in network speed, size and performance to be able to process arbitrarily large input data size. The size of the input data is only limited in practical terms by the electrical digital-to-analog converter memory. Hence, it is possible in principle to process images with 4K-resolution (4096×2160). Furthermore, by integrating 100 photonic convolution accelerators layers together, which is still far fewer than the typical numbers of processors in electronic chips, such as the 65536 processors integrated in the Google TPU [22] chip, the optical CNN would then be able to solve much more challenging image recognition functions, and with a vector computing speed of 100 × 11.3=1.130 Peta-OPS. Moreover, the network structure presented here is able to be trained online, since the optical Waveshaper used to weight the synapses can be reconfigured dynamically at speeds as fast as 500 milliseconds, or even potentially much faster with recently demonstrated optical integrated spectral shapers [147].

Even though our system had a reasonably large optical latency of over 100 nanosecond, mainly a result of the fibre spool, the latency does not affect the operational speed, which is a separate performance metric. Further, the latency of the system can be essentially eliminated to less than 200 picoseconds by using integrated highly dispersive elements such as customized chirped Bragg gratings or photonic crystals [148] or etalon based tunable dispersion compensators [149, 150]. Finally, existing nanofabrication technology can yield significantly higher levels of integration. The micro-comb source itself is already integrated and based on a CMOS compatible platform that is highly compatible with large-scale integration. Other components including the modulator, dispersive media, de-multiplexer optical spectral shaper, and photodetector have all been reported in integrated devices [147, 148, 151].

While optical neural networks are not yet at the level of performance as state-of-the-art electronic chips (>200 TOPs/s, scales with bit depth [13, 14, 15, 34]), our approach achieves operation speeds in the TeraOPs/s regime for the first time for optical networks. Further, there is enormous potential for scaling our systems through enhancing the spatial and wavelength dimensions and additional schemes such as using polarization. Both the convolutional accelerator and the CNN can be scaled in speed and processing power to enhance the parallelism using readily available off-the-shelf components. In the first instance, expanding the systems beyond the telecommunications C-band (1530-1570nm) to include the L-band (1570-1620nm) would yield a bandwidth of 90nm or 225 wavelengths (or channels) at a 50GHz spacing (0.4nm), versus the 90 wavelengths over 36nm in the C-band used here. These are both mainstream telecommunications bands for which there exists a tremendous amount of commercially available components and systems, including L-band EDFAs, Waveshapers, and many other components. Further, in the mainstream telecommunications bands (C+L) polarization sensitive components and devices are also available, meaning that taking advantage of polarization would yield an additional factor of 2x. Finally, spatial-division multiplexing, readily achievable using wavelength separation with either the Waveshaper or even just simple passive devices such as comb interleavers and passive filters, can offer, almost unlimited scalability, subject only to power/noise and scaling issues (cost, footprint, energy etc). Multiplying the system by a factor of at least 10, by using 10 parallel spatial paths, in principle is straightforward with existing components.

For the convolutional vector accelerator, operating with 3×3 kernels, and making use of polarization, the computational speed would be 2 × 2 × 9 × 62.9 = 2.26 TeraOPs/s per kernel. Making use of the C+L bands would produce 225 wavelengths at a 50GHz spacing, which would in turn allow 25 kernels, resulting in a processing speed of 25 × 2.26 = 56.6 TeraOPs/s. Using 10 spatial dimensions (through the Waveshaper) would enhance this to 0.57 PetaOPs/s.

The scale of the fully connected layer also has the potential to be significantly and readily increased with existing off-the-shelf technology. Since the number of neurons relies on spatial-division parallelism, or multiplexing, this number is, in principle, unlimited –subject only to tradeoffs in the OSNR (optical signal to noise ratio). By increasing the number of spatial paths (each with individual spectral shaping via more powerful WaveShapers and separate photo-detection), the number of neurons can be increased arbitrarily with existing instrumentation (subject to the SNR as mentioned). The number of synapses can also be significantly boosted though both wavelength and spatial division multiplexing. Making use of the full C+L band, supporting over 225 50GHz-spaced or 450 25GHz-spaced wavelength channels, and by exploiting dual polarization modes, the wavelength-division parallelism, and hence number of synapses per neuron could reach 225×2=450 (or even 900 at a 25GHz spacing, with tradeoffs in modulation rate). Further, even introducing a minimal number of additional spatial paths for each neuron (3 spatial paths, for example), the total number of synapses for 10 neurons @ 50GHz can reach 225 × 2 × 3 × 10 = 13,500 synapses in total.

Beyond this, a wider spectral region can readily be employed, although beyond the C+L bands, each has some challenges associated with it. Using the S+C+L telecommunications bands (1460-1620nm) would yield over 20THz in bandwidth. The telecommunications S-band (1460-1530nm), although less widely used than the mainstream C and L bands, is still practical with wideband optical devices available including semiconductor and Raman amplifiers. This would yield a total wavelength range of 160nm, equating to 400 channels at 50GHz spacing. Figure 11 shows a fully connected layer using the full C+L+S bands along with polarization, 3 spatial dimensions and 10 neurons, yielding 405 (wavelengths) × 2 (polarizations) × 3 (spatial paths) × 10 neurons = 24,300 synapses. Figure 12 shows the vector convolutional

accelerator, using the C+L+S bands (with 405 wavelengths) as well as 10 parallel spatial paths, and exploiting polarization. This would yield a speed of 62.9 Giga-Baud × 405 × 2 × 2 × 10 = 1.019 PetaOPs/s (or "POPs/s"). In this case the wavelengths would be distributed over 45 kernels each at 3×3 in size (so that 405=45×9).

Finally, in the long term, the full telecommunications bands including the O-band (1260 nm to 1360 nm) and even the E-band (water absorption band: 1360nm-1460nm) could be exploited, resulting in a total optical bandwidth of 1260nm-1620nm = 360nm, or 900 channels. Using the same arguments as above, extending the network to 50 neurons (which is feasible since only spatial multiplexing is used for this) the CNN could be expanded to yield 900 (wavelengths) × 2 (polarizations) × 3 (spatial paths) × 50 neurons = 270,000 synapses.

Note that in terms of optical bandwidth, micro-combs themselves are not a limiting factor – they have demonstrated full octave-spanning spectra – and more – from a single device, including from the near and mid-infrared [152-158] down to the visible region. One of the more restrictive components is the optical amplifier. While both C and L band amplifiers are in widespread use in installed optical fibre networks, they do not operate in any other band. Raman amplifiers are extremely flexible and versatile in wavelength and so can potentially operate in any of the telecommunications bands. SOAs as well are quite versatile with devices available in the O and S bands. While the Waveshaper has commercially developed for the C and L bands, the fundamental technology behind it (liquid crystal on silicon – LCOS) is capable of supporting operation in any of the telecommunications bands. The same holds true for most of the other components such as modulators, detectors etc. – while the commercially available components are generally designed for operation in the C and L bands, there is nothing fundamental in producing devices designed for the other bands – it is mostly a question of cost and scale.

In terms of the microcomb structure, the tradeoffs between comb FSR spacing and baud-rate are subject to the total available optical bandwidth, and are very similar to tradeoffs for ultrahigh bandwidth optical data communications. The computing speed of the accelerator is fundamentally determined by the available optical bandwidth. Within a certain optical band, the number of comb lines is inversely proportional to the FSR (i.e., the modulation rate). As long as the modulation rate matches the Nyquist bandwidth (half of the comb spacing), the network can be flexibly tailored to specific applications without sacrificing speed. In the case of the 49GHz microcombs studied here, as long as the optical band is sufficiently used, (i.e., the comb covers the full band and the modulation bandwidth (~24.5 GHz) matches with the FSR (~49GHz), the Nyquist bandwidth is ~ 24.5 GHz)), the computing speed does not vary dramatically with the number of comb lines or the FSR. So far, integrated microcombs feature FSRs ranging from 20 GHz to 1 THz, offering many options to choose from in terms of Baud-rate versus number of wavelengths. Having said this, we note that even for optical communications, this issue (the optimum channel spacing vs baud rate and modulation format) is still in fact an open question to a degree. Indeed, the exact optimum between the number of comb lines and the modulation rate is a function of the specific requirements for a given application. For applications that do not require a large number of kernel weights (wavelengths), a large FSR (modulation bandwidth) should be employed to make more extensive use of the optical band and achieve a high computing speed. While for those requiring a large number of kernel weights, a small FSR would be more favorable towards offering sufficient wavelengths.

Note that the preceding discussion does not address the issue of extending the CNN to much deeper levels. The electronic functions required for this have already been performed in this work, and include pooling, re-sampling, and re-timing. Further, some if not all of these can be realized all-optically. The pooling function can be implemented via the convolution accelerator with an averaging kernel (with all kernel weights set to be equal), followed with down-sampling to reduce the data scale. The reduction in speed of the convolutional accelerator when used for matrix processing, brought about by the overhead associated with flattening the matrix into a vector, is outlined in detail in [14], along with an example of a system architecture designed to eliminate this overhead for the case of an accelerator operating with 3×3 kernels, and in the process generating a symmetric convolution. We note that this is almost never an issue, however, and that asymmetric convolutions are the norm.

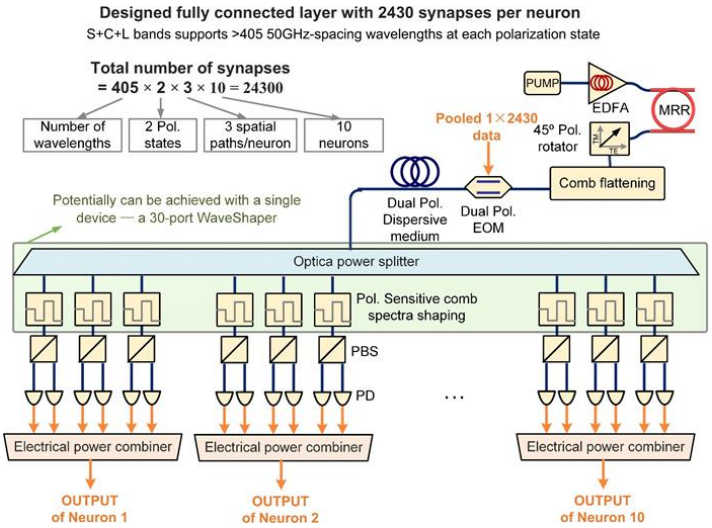

Figure 11. Designed expanded fully connected layer with 3 additional spatial dimensions and 10 neurons, making use of polarization multiplexing. Pump: continues-wave pump laser. EDFA: erbium doped fibre amplifier. MRR: micro-ring resonator. EOM: electro-optical Mach-Zehnder modulator. PBS: polarization beam splitter. PD: photodetector.

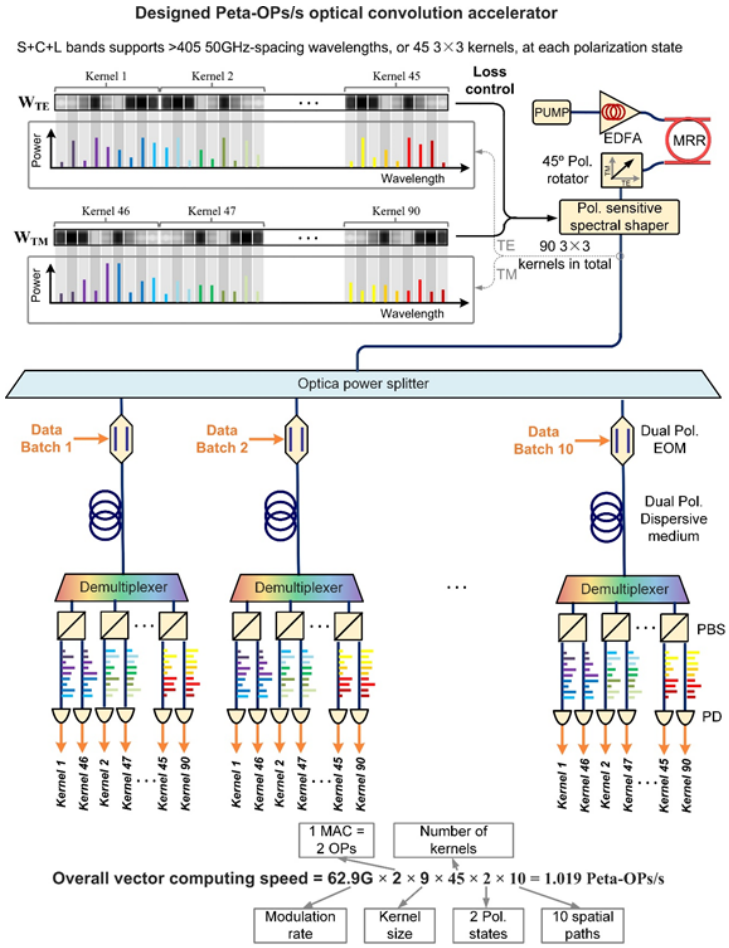

Figure 12. Designed scaled convolutional accelerator over the C+L+S bands, with spatial and polarization multiplexing. The 405 available wavelengths (on a 50GHz grid) would be split into 45 kernels each 3×3 in size.

5. CONCLUSION

We report an ultra-high speed universal optical convolutional accelerator that operates at speeds beyond 10 Tera-Operations per second (10 TOPS). It performs convolutions of ultra-large scale images at 250,000 pixels, with a digital resolution of 8-bits, and simultaneously for 10 kernels. This scale is more than adequate enough for facial image recognition tasks. We then employ the identical hardware to combine and form a sequential deep optical CNN with 10 output neurons. We successfully achieverecognition of the full 10 handwritten digits, each image being 900 pixels in size, achieving an accuracy of 88% which is very close to the theoretically calculated accuracy of 90% for our system. Our approach is trainable and scalable to much more complex networks to perform much more demanding tasks for applications such as real-time video recognition for unmanned autonomous vehicles.

REFERENCES

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature **521**, 436–444 (2015).
2. Schalkoff, R. J. Pattern recognition. Wiley Encyclopedia of Computer Science and Engineering (2007).
3. Mnih, V. et al. Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015).
4. Silver, D. et al. Mastering the game of Go without human knowledge. Nature **550**, 354 (2017).
5. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Commun Acm 60, 84-90 (2017).
6. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. Nature **577**, 641–646 (2020).
7. Lawrence, S. et al., "Face recognition: A convolutional neural-network approach", IEEE transactions on neural networks 8, 98-113 (1997).
8. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. Nature Photonics **11**, 441 (2018).
9. Larger, L. et al. High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification. Phys. Rev. X **7**, 011015 (2017).
10. Peng, H., Nahmias, M. A., Lima, T. F. d., Tait, A. N. & Shastri, B. J. Neuromorphic Photonic Integrated Circuits. IEEE Journal of Selected Topics in Quantum Electronics **24**, 1-15, doi:10.1109/JSTQE.2018.2840448 (2018).
11. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. Science **361**, 1004-1008 (2018).
12. Feldmann, J. et al., "All-optical spiking neurosynaptic networks with self-learning capabilities", Nature **569**, 208-214 (2019).
13. X. Xu et al, "Photonic perceptron based on a Kerr microcomb for scalable high speed optical neural networks", Laser and Photonics Reviews, vol. 14, no. 8, 2000070 (2020). DOI: 10.1002/lpor.202000070.
14. X. Xu, et al., "11 TOPs photonic convolutional accelerator for optical neural networks", Nature **589**, 44-51 (2021).
15. Feldmann, J. *et al.*, "Parallel convolutional processing using an integrated photonic tensor core", Nature **589**, 52-58 (2021).
16. T. Zhou et al, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit", Nature Photonics Vol 15, (5) 367 (2021).
17. B. J. Shastri *et al.,* "Photonics for artificial intelligence and neuromorphic computing", Nature Photonics **15,** (2) 102-114 (2021).
18. H. Wu, and Q. Dai, "Artificial intelligence accelerated by light", Nature **589**, 25-26 (2021).
19. G. Wetzstein *et al.,* "Inference in artificial intelligence with deep optics and photonics", Nature **588** (7836), 39-47 (2020).
20. Ambrogio, S. *et al.,* "Equivalent-accuracy accelerated neural-network training using analogue memory", *Nature* **558**, 60 (2018).
21. Esser, S. K. et al., "Convolutional networks for fast, energy-efficient neuromorphic computing", Proc. of the National Academy of Sciences **113**, 11441 (2016).
22. Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. Nature **538**, 471–476 (2016).
23. Miller, D. A. B. Attojoule Optoelectronics for Low-Energy Information Processing and Communications. Journal of Lightwave Technology **35**, 346-396 (2017).
24. Appeltant, L. et al. Information processing using a single dynamical node as complex system. Nature Communications **2**, 468 (2011).
25. Chang, J., Sitzmann, V., Dun, X., Heidrich, W. & Wetzstein, G. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Scientific Reports 8 (2018).
26. Vandoorne, K. et al., "Experimental demonstration of reservoir computing on a silicon photonics chip", Nature Communications **5**, 3541 (2014).
27. Brunner, D. et al., "Parallel photonic information processing at gigabyte per second data rates using transient states", Nature Communications **4**, 1364 (2013).
28. Tait, A. N. et al., "Demonstration of WDM weighted addition for principal component analysis", Optics Express **23**, 12758-12765 (2015).
29. Pasquazi, A. et al. Micro-combs: a novel generation of optical sources. Physics Reports **729**, 1-81 (2018).
30. Moss, D. J. et al., "New CMOS-compatible platforms based on silicon nitride and Hydex for nonlinear optics", Nature photonics **7**, 597 (2013).
31. Kippenberg, T. J., Gaeta, A. L., Lipson, M. & Gorodetsky, M. L. Dissipative Kerr solitons in optical microresonators. Science **361**, 567 (2018).
32. Savchenkov, A. A. et al. Tunable optical frequency comb with a crystalline whispering gallery mode resonator. Physics Review Letters **101**, 093902 (2008).
33. Spencer, D. T. et al. An optical-frequency synthesizer using integrated photonics. Nature **557**, 81-85 (2018).
34. Marin-Palomo, P. et al. Microresonator-based solitons for massively parallel coherent optical communications. Nature **546**, 274 (2017).
35. B. Corcoran, et al., "Ultra-dense optical data transmission over standard fiber with a single chip source", Nature Communications, vol. 11, Article:2568, 2020.
36. Kues, M. et al. Quantum optical microcombs. Nature Photonics **13**, (3) 170-179 (2019). doi:10.1038/s41566-019-0363-0
37. C. Reimer, L. Caspani, M. Clerici, et al., "Integrated frequency comb source of heralded single photons," Optics Express, vol. 22, no. 6, pp. 6535-6546, 2014.
38. C. Reimer, et al., "Cross-polarized photon-pair generation and bi-chromatically pumped optical parametric oscillation on a chip", Nature Communications, vol. 6, Article 8236, 2015. DOI: 10.1038/ncomms9236.
39. L. Caspani, C. Reimer, M. Kues, et al., "Multifrequency sources of quantum correlated photon pairs on-chip: a path toward integrated Quantum Frequency Combs," Nanophotonics, vol. 5, no. 2, pp. 351-362, 2016.

40. C. Reimer et al., "Generation of multiphoton entangled quantum states by means of integrated frequency combs," Science, vol. 351, no. 6278, pp. 1176-1180, 2016.
41. M. Kues, et al., "On-chip generation of high-dimensional entangled quantum states and their coherent control", Nature, vol. 546, no. 7660, pp. 622-626, 2017.
42. P. Roztocki et al., "Practical system for the generation of pulsed quantum frequency combs," Optics Express, vol. 25, no. 16, pp. 18940-18949, 2017.
43. Y. Zhang, et al., "Induced photon correlations through superposition of two four-wave mixing processes in integrated cavities", Laser and Photonics Reviews, vol. 14, no. 7, pp. 2000128, 2020. DOI: 10.1002/lpor.202000128
44. C. Reimer, et al.,"High-dimensional one-way quantum processing implemented on d-level cluster states", Nature Physics, vol. 15, no.2, pp. 148–153, 2019.
45. Stern, B., Ji, X., Okawachi, Y., Gaeta, A. L. & Lipson, M. Battery-operated integrated frequency comb generator. Nature **562**, 401 (2018).
46. H. Bao, et al., Laser cavity-soliton microcombs, Nature Photonics, vol. 13, no. 6, pp. 384-389, Jun. 2019.
47. Lugiato, L. A., Prati, F. & Brambilla, M. Nonlinear Optical Systems, (Cambridge University Press, 2015).
48. Cole, D. C., Lamb, E. S., Del'Haye, P., Diddams, S. A. & Papp, S. B. Soliton crystals in Kerr resonators. Nature Photonics **11**, 671 (2017).
49. Wang, W., et al.., Robust soliton crystals in a thermally controlled microresonator, Opt. Lett., 43, 2002 (2018).
50. Bao, C., et al., Direct soliton generation in microresonators, Opt. Lett, **42**, 2519 (2017).
51. M.Ferrera et al., "CMOS compatible integrated all-optical RF spectrum analyzer", Optics Express, vol. 22, no. 18, 21488 - 21498 (2014).
52. A. Pasquazi, et al., "Sub-picosecond phase-sensitive optical pulse characterization on a chip", Nature Photonics, vol. 5, no. 10, pp. 618-623 (2011).
53. M. Kues, et al., "Passively modelocked laser with an ultra-narrow spectral width", Nature Photonics, vol. 11, no. 3, pp. 159, 2017.
54. L. Razzari, et al., "CMOS-compatible integrated optical hyper-parametric oscillator," Nature Photonics, vol. 4, no. 1, pp. 41-45, 2010.
55. M. Ferrera, et al., "Low-power continuous-wave nonlinear optics in doped silica glass integrated waveguide structures," Nature Photonics, vol. 2, no. 12, pp. 737-740, 2008.
56. M.Ferrera et al."On-Chip ultra-fast 1st and 2nd order CMOS compatible all-optical integration", Opt. Express, vol. 19, (23)pp. 23153-23161 (2011).
57. D. Duchesne, M. Peccianti, M. R. E. Lamont, et al., "Supercontinuum generation in a high index doped silica glass spiral waveguide," Optics Express, vol. 18, no, 2, pp. 923-930, 2010.
58. H Bao, L Olivieri, M Rowley, ST Chu, BE Little, R Morandotti, DJ Moss, ... "Turing patterns in a fiber laser with a nested microresonator: Robust and controllable microcomb generation", Physical Review Research **2** (2), 023395 (2020).
59. M. Ferrera, et al., "On-chip CMOS-compatible all-optical integrator", Nature Communications, vol. 1, Article 29, 2010.
60. A. Pasquazi, et al., "All-optical wavelength conversion in an integrated ring resonator," Optics Express, vol. 18, no. 4, pp. 3858-3863, 2010.
61. A. Pasquazi, Y. Park, J. Azana, et al., "Efficient wavelength conversion and net parametric gain via Four Wave Mixing in a high index doped silica waveguide," Optics Express, vol. 18, no. 8, pp. 7634-7641, 2010.
62. M. Peccianti, M. Ferrera, L. Razzari, et al., "Subpicosecond optical pulse compression via an integrated nonlinear chirper," Optics Express, vol. 18, no. 8, pp. 7625-7633, 2010.
63. Little, B. E. et al., "Very high-order microring resonator filters for WDM applications", IEEE Photonics Technol. Lett. **16**, 2263–2265 (2004).
64. M. Ferrera et al., "Low Power CW Parametric Mixing in a Low Dispersion High Index Doped Silica Glass Micro-Ring Resonator with Q-factor > 1 Million", Optics Express, vol.17, no. 16, pp. 14098–14103 (2009).
65. M. Peccianti, et al., "Demonstration of an ultrafast nonlinear microcavity modelocked laser", Nature Communications, vol. 3, pp. 765, 2012.
66. A. Pasquazi, et al., "Self-locked optical parametric oscillation in a CMOS compatible microring resonator: a route to robust optical frequency comb generation on a chip," Optics Express, vol. 21, no. 11, pp. 13333-13341, 2013.
67. A. Pasquazi, et al., "Stable, dual mode, high repetition rate mode-locked laser based on a microring resonator," Optics Express, vol. 20, no. 24, pp. 27355-27362, 2012.
68. Wu, J. *et al.* RF Photonics: An Optical Microcombs' Perspective. IEEE Journal of Selected Topics in Quantum Electronics Vol. **24**, 6101020, 1-20 (2018).
69. Xu, X., et al., Photonic microwave true time delays for phased array antennas using a 49 GHz FSR integrated micro-comb source, *Photonics Research*, **6**, B30-B36 (2018).
70. T. G. Nguyen *et al.,* "Integrated frequency comb source-based Hilbert transformer for wideband microwave photonic phase analysis," *Opt. Express,* vol. 23, no. 17, pp. 22087-22097, Aug. 2015.
71. X. Xu, J. Wu, M. Shoeiby, T. G. Nguyen, S. T. Chu, B. E. Little, R. Morandotti, A. Mitchell, and D. J. Moss, "Reconfigurable broadband microwave photonic intensity differentiator based on an integrated optical frequency comb source," *APL Photonics,* vol. 2, no. 9, 096104, Sep. 2017.
72. X. Xu, M. Tan, J. Wu, R. Morandotti, A. Mitchell, and D. J. Moss, "Microcomb-based photonic RF signal processing", *IEEE Photonics Technology Letters,* vol. 31 no. 23 1854-1857, 2019.
73. X. Xu, *et al.*, "Broadband RF channelizer based on an integrated optical frequency Kerr comb source," *Journal of Lightwave Technology,* vol. 36, no. 19, pp. 4519-4526, 2018.
74. X. Xu, *et al.*, "Continuously tunable orthogonally polarized RF optical single sideband generator based on micro-ring resonators," *Journal of Optics,* vol. 20, no. 11, 115701. 2018.
75. X. Xu, *et al.*, "Orthogonally polarized RF optical single sideband generation and dual-channel equalization based on an integrated microring resonator," *Journal of Lightwave Technology,* vol. 36, no. 20, pp. 4808-4818. 2018.

76. M.Tan, X. Xu, J. Wu, B. Corcoran, A. Boes, T. G. Nguyen, Sai T. Chu, B. E. Little, R.Morandotti, A. Mitchell, and D. J. Moss, "Integral order photonic RF signal processors based on a soliton crystal micro-comb source", IOP Journal of Optics vol. 23 (11) 125701 (2021).
77. X. Xu, *et al.*, "Advanced adaptive photonic RF filters with 80 taps based on an integrated optical micro-comb source," *Journal of Lightwave Technology*, vol. 37, no. 4, pp. 1288-1295, 2019.
78. X. Xu, *et al.*, Broadband microwave frequency conversion based on an integrated optical micro-comb source", *Journal of Lightwave Technology*, vol. 38 no. 2, pp. 332-338, 2020.
79. M. Tan, *et al.*, "Photonic RF and microwave filters based on 49GHz and 200GHz Kerr microcombs", *Optics Comm*. vol. 465,125563, Feb. 22. 2020.
80. X. Xu, *et al.*, "Broadband photonic RF channelizer with 90 channels based on a soliton crystal microcomb", *Journal of Lightwave Technology*, Vol. 38, no. 18, pp. 5116 - 5121, 2020. doi: 10.1109/JLT.2020.2997699.
81. X. Xu, *et al.*, "Photonic RF and microwave integrator with soliton crystal microcombs", *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3582-3586, 2020. DOI:10.1109/TCSII.2020.2995682.
82. X. Xu, *et al.*, "Photonic RF phase-encoded signal generation with a microcomb source", *J. Lightwave Technology*, vol. 38, no. 7, 1722-1727, 2020.
83. X. Xu, *et al.*, "High performance RF filters via bandwidth scaling with Kerr micro-combs," *APL Photonics*, vol. 4, no. 2, pp. 026102. 2019.
84. M. Tan, *et al.*, "Microwave and RF photonic fractional Hilbert transformer based on a 50 GHz Kerr micro-comb", *Journal of Lightwave Technology*, vol. 37, no. 24, pp. 6097 – 6104, 2019.
85. M. Tan, *et al.*, "RF and microwave fractional differentiator based on photonics", *IEEE Transactions on Circuits and Systems: Express Briefs*, vol. 67, no.11, pp. 2767-2771, 2020. DOI:10.1109/TCSII.2020.2965158.
86. M. Tan, *et al*., "Photonic RF arbitrary waveform generator based on a soliton crystal micro-comb source", Journal of Lightwave Technology, vol. 38, no. 22, pp. 6221-6226, Oct 22. 2020. DOI: 10.1109/JLT.2020.3009655.
87. M. Tan, X. Xu, J. Wu, R. Morandotti, A. Mitchell, and D. J. Moss, "RF and microwave high bandwidth signal processing based on Kerr Micro-combs", Advances in Physics X, VOL. 6, NO. 1, 1838946 (2021). DOI:10.1080/23746149.2020.1838946.
88. X. Xu, et al., "Advanced RF and microwave functions based on an integrated optical frequency comb source," Opt. Express, vol. 26 (3) 2569 2018.
89. Kues, M. et al. Quantum optical microcombs. Nature Photonics **13**, (3) 170-179 (2019). doi:10.1038/s41566-019-0363-0
90. P.Roztocki et al., "Complex quantum state generation and coherent control based on integrated frequency combs", Journal of Lightwave Technology **37** (2) 338-347 (2019).
91. S. Sciara et al., "Generation and Processing of Complex Photon States with Quantum Frequency Combs", IEEE Photonics Technology Letters **31** (23) 1862-1865 (2019). DOI: 10.1109/LPT.2019.2944564.
92. L. Caspani, C. Reimer, M. Kues, et al., "Multifrequency sources of quantum correlated photon pairs on-chip: a path toward integrated Quantum Frequency Combs," Nanophotonics, vol. 5, no. 2, pp. 351-362, 2016.
93. C. Reimer et al., "Generation of multiphoton entangled quantum states by means of integrated frequency combs," Science, vol. 351, no. 6278, pp. 1176-1180, 2016.
94. M. Kues, et al., "On-chip generation of high-dimensional entangled quantum states and their coherent control", Nature, vol. 546, no. 7660, pp. 622-626, 2017.
95. P. Roztocki et al., "Practical system for the generation of pulsed quantum frequency combs," Optics Express, vol. 25, no.16, 18940-18949, 2017.
96. Y. Zhang, et al., "Induced photon correlations through superposition of two four-wave mixing processes in integrated cavities", Laser and Photonics Reviews, vol. 14, no. 7, pp. 2000128, 2020. DOI: 10.1002/lpor.202000128
97. C. Reimer, et al.,"High-dimensional one-way quantum processing implemented on d-level cluster states", Nature Physics, vol. 15 (2) 148 (2019).
98. H. Bao, et al., Laser cavity-soliton microcombs, Nature Photonics, vol. 13, no. 6, pp. 384-389, Jun. 2019.
99. Bao, C., et al., Direct soliton generation in microresonators, Opt. Lett, **42**, 2519 (2017).
100. M.Ferrera et al., "CMOS compatible integrated all-optical RF spectrum analyzer", Optics Express, vol. 22, (18) 21488 (2014).
101. A. Pasquazi, et al., "Sub-picosecond phase-sensitive optical pulse characterization on a chip", Nature Photonics, vol. 5, no. 10, pp. 618-623 (2011).
102. M. Kues, et al., "Passively modelocked laser with an ultra-narrow spectral width", Nature Photonics, vol. 11, no. 3, pp. 159, 2017.
103. L. Razzari, et al., "CMOS-compatible integrated optical hyper-parametric oscillator," Nature Photonics, vol. 4, no. 1, 41-45, 2010.
104. M. Ferrera, et al., "Low-power continuous-wave nonlinear optics in doped silica glass integrated waveguide structures," Nature Photonics, vol. 2, no. 12, pp. 737-740, 2008.
105. M.Ferrera et al."On-Chip ultra-fast 1st and 2nd order CMOS compatible all-optical integration", Opt. Express, vol. 19, (23)pp. 23153-23161 (2011).
106. D. Duchesne, M. Peccianti, M. R. E. Lamont, et al., "Supercontinuum generation in a high index doped silica glass spiral waveguide," Optics Express, vol. 18, no, 2, pp. 923-930, 2010.
107. H Bao, L Olivieri, M Rowley, ST Chu, BE Little, R Morandotti, DJ Moss, ... "Turing patterns in a fiber laser with a nested microresonator: Robust and controllable microcomb generation", Physical Review Research **2** (2), 023395 (2020).
108. M. Ferrera, et al., "On-chip CMOS-compatible all-optical integrator", Nature Communications, vol. 1, Article 29, 2010.
109. A. Pasquazi, et al., "All-optical wavelength conversion in an integrated ring resonator," Optics Express, vol. 18 (4) 3858 (2010).
110. A. Pasquazi, Y. Park, J. Azana, et al., "Efficient wavelength conversion and net parametric gain via Four Wave Mixing in a high index doped silica waveguide," Optics Express, vol. 18, no. 8, pp. 7634-7641, 2010.
111. M. Peccianti, M. Ferrera, L. Razzari, et al., "Subpicosecond optical pulse compression via an integrated nonlinear chirper," Optics Express, vol. 18, no. 8, pp. 7625-7633, 2010.
112. Little, B. E. et al., "Very high-order microring resonator filters for WDM applications", IEEE Phot. Technol. Lett. **16**, 2263(2004).

113. M. Ferrera et al., “Low Power CW Parametric Mixing in a Low Dispersion High Index Doped Silica Glass Micro-Ring Resonator with Q-factor > 1 Million”, Optics Express, vol.17, no. 16, pp. 14098–14103 (2009).
114. M. Peccianti, et al., “Demonstration of an ultrafast nonlinear microcavity modelocked laser”, Nature Comm., vol. 3, 765, 2012.
115. A. Pasquazi, et al., “Self-locked optical parametric oscillation in a CMOS compatible microring resonator: a route to robust optical frequency comb generation on a chip,” Optics Express, vol. 21, no. 11, pp. 13333-13341, 2013.
116. A. Pasquazi, et al., “Stable, dual mode, high repetition rate mode-locked laser based on a microring resonator,” Optics Express, vol. 20, no. 24, pp. 27355-27362, 2012.
117. Xu, X., et al., Photonic microwave true time delays for phased array antennas using a 49 GHz FSR integrated micro-comb source, *Photonics Research*, **6**, B30-B36 (2018).
118. X. Xu, M. Tan, J. Wu, R. Morandotti, A. Mitchell, and D. J. Moss, “Microcomb-based photonic RF signal processing”, *IEEE Photonics Technology Letters*, vol. 31 no. 23 1854-1857, 2019.
119. M. Tan et al, “Orthogonally polarized Photonic Radio Frequency single sideband generation with integrated micro-ring resonators”, IOP Journal of Semiconductors, Vol. **42** (4), 041305 (2021). DOI: 10.1088/1674-4926/42/4/041305.
120. Xu, *et al.*, “Advanced adaptive photonic RF filters with 80 taps based on an integrated optical micro-comb source,” *Journal of Lightwave Technology*, vol. 37, no. 4, pp. 1288-1295 (2019).
121. X. Xu, *et al.,* Broadband microwave frequency conversion based on an integrated optical micro-comb source”, *Journal of Lightwave Technology*, vol. 38 no. 2, pp. 332-338, 2020.
122. M. Tan, *et al.,* “Photonic RF and microwave filters based on 49GHz and 200GHz Kerr microcombs”, *Optics Comm*. vol. 465,125563, Feb. 22. 2020.
123. X. Xu, *et al.,* “Broadband photonic RF channelizer with 90 channels based on a soliton crystal microcomb”, *Journal of Lightwave Technology*, Vol. 38, no. 18, pp. 5116 - 5121, 2020. doi: 10.1109/JLT.2020.2997699.
124. X. Xu, *et al.,* “Photonic RF and microwave integrator with soliton crystal microcombs”, *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3582-3586, 2020. DOI:10.1109/TCSII.2020.2995682.
125. X. Xu, *et al.*, “High performance RF filters via bandwidth scaling with Kerr micro-combs,” *APL Photonics*, vol. 4 (2) 026102. 2019.
126. M. Tan, *et al.,* “Microwave and RF photonic fractional Hilbert transformer based on a 50 GHz Kerr micro-comb”, *Journal of Lightwave Technology*, vol. 37, no. 24, pp. 6097 – 6104, 2019.
127. M. Tan, *et al.,* “RF and microwave fractional differentiator based on photonics”, *IEEE Transactions on Circuits and Systems: Express Briefs*, vol. 67, no.11, pp. 2767-2771, 2020. DOI:10.1109/TCSII.2020.2965158.
128. M. Tan, *et al*., “Photonic RF arbitrary waveform generator based on a soliton crystal micro-comb source”, Journal of Lightwave Technology, vol. 38, no. 22, pp. 6221-6226 (2020). DOI: 10.1109/JLT.2020.3009655.
129. M. Tan, X. Xu, J. Wu, R. Morandotti, A. Mitchell, and D. J. Moss, “RF and microwave high bandwidth signal processing based on Kerr Micro-combs”, Advances in Physics X, VOL. 6, NO. 1, 1838946 (2021). DOI:10.1080/23746149.2020.1838946.
130. X. Xu, et al., “Advanced RF and microwave functions based on an integrated optical frequency comb source,” Opt. Express, vol. 26 (3) 2569 (2018).
131. M. Tan, X. Xu, J. Wu, B. Corcoran, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, R.Morandotti, A. Lowery, A. Mitchell, and D. J. Moss, “"Highly Versatile Broadband RF Photonic Fractional Hilbert Transformer Based on a Kerr Soliton Crystal Microcomb”, Journal of Lightwave Technology vol. 39 (24) 7581-7587 (2021).
132. Bao, C., et al., Direct soliton generation in microresonators, Opt. Lett, 42, 2519 (2017).
133. Yuning Zhang, Yang Qu, Jiayang Wu, Linnan Jia, Yunyi Yang, Xingyuan Xu, Baohua Jia, and David J. Moss, “Enhanced Kerr nonlinearity and nonlinear figure of merit in silicon nanowires integrated with 2D graphene oxide films”, ACS Applied Materials and Interfaces, Vol. 12 (29) 33094−33103 (2020). DOI:10.1021/acsami.0c07852
134. D.Moss, “11 Tera-FLOP/s photonic convolutional accelerator and deep learning optical neural networks”, Research Square, (2021). DOI: https://doi.org/10.21203/rs.3.rs-493347/v1.
135. Moss, David (2020): “11.0 Tera-FLOP/second photonic convolutional accelerator for deep learning optical neural networks”, TechRxiv. Preprint. (2020). https://doi.org/10.36227/techrxiv.13238423.v1
136. Xu, X.; Tan, M.; Corcoran, B.; Wu, J.; Boes, A.; Nguyen, T.; Chu, S.; Little, B.; Hicks, D.; Morandotti, R.; Mitchell, A.; Moss, D. “11 Tera-FLOP per Second Photonic Convolutional Accelerator for Deep Learning Optical Neural Networks”, Preprints 2020, 2020110420.
137. Moss, David (2020): “RF and microwave photonic high bandwidth signal processing based on Kerr micro-comb sources”, TechRxiv. (2020). Preprint. DOI:10.36227/techrxiv.12665609.v3
138. Yuning Zhang, Jiayang Wu, Yunyi Yang, Yang Qu, Linnan Jia, Tania Moein, Baohua Jia, David J. Moss, “Enhanced nonlinear optical figure-of-merit at 1550nm for silicon nanowires integrated with graphene oxide layered films”, Arxiv (2020). arXiv:2004.08043 [physics.optics]
139. Moss, David; Jia, Baohua; Wu, Jiayang; Zhang, Yuning; Yang, Yunyi; Jia, Linnan, Yang Qu, Tania Moein (2020): “Transforming silicon into a high performing integrated nonlinear photonics platform by integration with 2D graphene oxide films”, TechRxiv. (2020). Preprint. DOI:10.36227/techrxiv.12061809.v1.
140. A. Frigg, A. Boes, G. Ren, T.G. Nguyen, D.Y. Choi, S. Gees, D. Moss, A Mitchell, “Optical frequency comb generation with low temperature reactive sputtered silicon nitride waveguides”, APL Photonics, Vol. 5 (1), 011302 (2020).
141. T. Moein, D. Gailevičius, T. Katkus, S.H. Ng, S. Lundgaard, D.J. Moss, H. Kurt, Vygantas Mizeikis, Kęstutis Staliūnas, Mangirdas Malinauskas, Saulius Juodkazis, “Optically-thin broadband graphene-membrane photodetector”, Nanomaterials, Vol. 10 (3), 407 (2020).
142. M. Tan, X. Xu, J. Wu, A. Boes, B. Corcoran, T. G. Nguyen, S. T. Chu, B. E. Little, R. Morandotti, A. Mitchell, and D. J. Moss, “Advanced applications of Kerr mircocombs”, Paper 11775-1. SPIE 11775, Integrated Optics: Design, Devices, Systems and Applications VI, (EOO21) OO107-8, Proc 1177504 (18 April 2021); Integrated Optics Conference, SPIE Optics and Optoelectronics Symposium, Prague, Czech Republic. April 19 - 22 (2021), doi.org/10.1117/12.2588733.

143. Moss, David, "Microcombs for Ultrahigh Bandwidth Optical Data Transmission and Neural Networks." OSF Preprints. March 8. (2021). DOI:10.31219/osf.io/ne9wx.
144. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097-1105 (2012).
145. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**, 2278-2324 (1998).
146. Bishop, C. M. Neural networks for pattern recognition. (Oxford university press, 1995).
147. Metcalf, A. J. et al. Integrated line-by-line optical pulse shaper for high-fidelity and rapidly reconfigurable RF-filtering. Optics Express **24**, 23925-23940 (2016).
148. Sahin, E et al., "Large, scalable dispersion engineering using cladding-modulated Bragg gratings on a silicon chip", Applied Physics Letters **110**, 161113 (2017).
149. D. J. Moss et al.,"Tunable dispersion and dispersion slope compensators for 10Gb/s using all-pass multicavity etalons", IEEE Phot. Technology Letters, vol. 15, no. 5, 730-732 (2003).
150. L.M. Lunardi et al. ,"Tunable dispersion compensators based on multi-cavity all-pass etalons for 40Gb/s systems", J. Lightwave Technology, vol. 20, (12) 2136 (2002).
151. Wang, C. et al. Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. Nature **562**, 101 (2018).
152. A. Della Torre et al., "Mid-infrared supercontinuum generation in a low-loss germanium-on-silicon waveguide", APL Photonics Vol. 6, 016102 (2021); doi: 10.1063/5.0033070.
153. M. Sinobad, et al., "Mid-infrared supercontinuum generation in silicon-germanium all-normal dispersion waveguides", Optics Letters, Vol. 45 (18), 5008-5011 (2020). DOI: 10.1364/OL.402159.
154. M. Sinobad et al., "High coherence at f and 2f of a mid-infrared supercontinuum in a silicon germanium waveguide", IEEE Journal of Selected Topics in Quantum Electronics Vol. 26 (2) 8201008 (2020). DOI:10.1109/JSTQE.2019.2943358.
155. M. Sinobad et al., "Dispersion trimming for mid-infrared supercontinuum generation in a hybrid chalcogenide Si-Ge waveguide", Journal of the Optical Society of America B, Vol. 36 (2) A98-A104 (2019). DOI: 10.1364/JOSAB.36.000A98.
156. M. Sinobad et al., "High brightness mid-infrared octave spanning supercontinuum generation to 8.5μm in chip-based Si-Ge waveguides", Optica, Vol. 5 (4) 360-366 (2018). DOI:10.1364/OPTICA.5.000360.
157. L. Jin et al., Applied Physics Letters Photonics, Vol, 5 Article 056106, (2020). DOI:10.1063/5.0002941
158. L. Carletti et al., "Nonlinear optical properties of Si-Ge waveguides in the mid-infrared", Optics Express Vol. 23 (7) 8261–8271 (2015).