# Optical neuromorphic processing with Kerr microcombs: Scaling the network in size and speed to the PetaOp regime

**David Moss** ( dmoss@swin.edu.au )
Swinburne University of Technology

**Research Article**

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

**Posted Date:** January 17th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1266795/v1

# Abstract

Convolutional neural networks (CNNs), inspired by biological visual cortex systems, are a powerful category of artificial neural networks that can extract the hierarchical features of raw data to greatly reduce the network parametric complexity and enhance the predicting accuracy. They 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 offer the promise of dramatically accelerating computing speed to overcome the inherent bandwidth bottleneck of electronics. Here, we demonstrate a universal optical vector convolutional accelerator operating beyond 10 Tera-OPS (TOPS - operations per second), generating convolutions of images of 250,000 pixels with 8-bit resolution for 10 kernels simultaneously — enough for facial image recognition. We then use the same hardware to sequentially form a deep optical CNN with ten output neurons, achieving successful recognition of full 10 digits with 900 pixel handwritten digit images with 88% accuracy. Our results are based on simultaneously interleaving temporal, wavelength and spatial dimensions enabled by an integrated microcomb source. We show that this approach is scalable and trainable to much more complex networks for demanding applications such as unmanned vehicle and real-time video recognition.


# 1. Introduction

Artificial neural networks (ANNs) are collections of nodes with weighted connections that, with proper feedback to adjust the network parameters, can “learn” and perform complex operations for face recognition, speech translation, playing board games and medical diagnosis [1- 4]. While classic fully connected feedforward networks face challenges in processing extremely high-dimensional data, convolutional neural networks (CNNs), inspired by the (biological) behavior of the visual cortex system, can abstract the representations of input data in their raw form, and then predict their properties with both unprecedented accuracy and greatly reduced parametric complexity [5]. CNNs have been widely applied to computer vision, natural language processing and other areas [6, 7].

The capability of neural networks is dictated by the computing power of the underlying neuromorphic hardware. Optical neural networks (ONNs) [8-19] are promising candidates for next-generation neuromorphic computation, since they have the potential to overcome the bandwidth bottleneck of their electrical counterparts [6, 20-23] and achieve ultra-high computing speeds enabled by the >10 THz wide optical telecom band [8]. ONNs are attracting a great deal of attention with recent breakthroughs and reviews [13-19]. Operating in analog frameworks, they avoid the limitations imposed by the energy and time consumed during reading and storing data back and forth, known as the von Neumann bottleneck [20]. Significant progress has been made in highly parallel, high-speed and trainable ONNs [8-19, 24-28], including approaches that have the potential for full integration on a single photonic chip [8,12,14,15], in turn offering an ultra-high computational density. However, there remains opportunities for significant improvements in ONNs. Processing large-scale data, as needed for practical real-life computer vision tasks, remains challenging because they are primarily fully connected structures where their input scale is

determined solely by hardware parallelism. This leads to tradeoffs between the network scale and footprint. Moreover, ONNs have not achieved the extreme computing speeds that analog photonics is capable of.

Here, we demonstrate an optical convolution accelerator to process and compress large-scale data. Through interleaving wavelength, temporal, and spatial dimensions using an integrated Kerr frequency comb, or microcomb [29 – 143], we achieve a vector computing speed as high as 11.322 TOPS and use it to process 250,000 pixel images with 10 convolution kernels at 3.8 TOPs. The convolution accelerator is fully and dynamically reconfigurable, and scalable, and can serve as both a convolutional accelerator front-end with multiple and simultaneous parallel kernels, as well as forming an optically deep CNN with fully connected neurons, with the same hardware. We demonstrate a CNN and successfully apply it to the recognition of full ten digit (0-9) handwritten images, achieving an accuracy of 88%. We then present further architectures to scale the network in speed to the Peta-OP regime as well as to over 24,000 synapses, by using the full S,C,L telecommunications wavelength bands.

Our optical neural network represents a major step towards realizing monolithically integrated ONNs and is enabled by our use of an integrated microcomb chip. Moreover, our accelerator scheme is stand alone and universal — fully compatible with either electrical or optical interfaces. Hence, it can serve as a universal ultrahigh bandwidth data compressing front end for any neuromorphic hardware — either optical or electronic — making massive-data machine learning for real-time, ultrahigh bandwidth data possible.

# 2. Principle Of Operation

Figure 1 shows the operation principle of the photonic convolutional accelerator (CA), featuring high-speed electrical signal input and output data ports, while Figure 2 shows a detailed experimental configuration. The data vector input X is serially encoded with the intensity of temporal symbols in an electrical waveform at a symbol rate $1/\tau$ (baud), where $\tau$ is the symbol period. The convolution kernel is likewise represented by a weight vector W of length R that is used to encode the optical power of the microcomb lines by spectral shaping with a Waveshaper. The temporal waveform X is then multi-cast onto the kernel wavelength channels via electro-optical modulation, generating the replicas weighted by W. Next the optical waveform is transmitted through a dispersive delay with a delay step between adjacent wavelength channels equal to the symbol duration of X, thus achieving time and wavelength interleaving. Finally, the delayed and weighted replicas are summed via high speed photodetection so that each time slot yields a convolution between X and W for a given convolution window, or receptive field. Thus, the convolution window effectively slides at the modulation speed matching the baud rate of X. Each output symbol is the result of R multiply-and-accumulate operations, with the computing speed given by $2R/\tau$ OPS. Since the speed of this process scales with both the baud rate and number of wavelengths, it can be dramatically boosted into the TOP regime by using the massively parallel wavelength channels of a microcomb. Further, the input data X length is unlimited - the convolution accelerator can process arbitrarily large- scale data, limited only by the electronics. Likewise, the kernels number and length are arbitrary, limited only by the number of wavelengths. We achieve simultaneous

convolution of multiple kernels by adding additional sub-bands of R wavelengths for each kernel. Following multicasting and dispersive delay, the sub-bands (kernels) are demultiplexed and detected separately with high speed photodetectors, generating a separate electronic waveform for each kernel.

While the convolutional accelerator typically processes vectors, it can operate on matrices for image processing by flattening the matrix into a vector. The precise way that this is done determines both the sliding convolution window's stride and the equivalent matrix computing speed. Our flattening method sets the receptive field (convolution slot) to slide with a horizontal stride of unity (ie., every matrix input element has a corresponding convolution output) and a vertical stride that scales with the size of the convolutional kernel. The larger vertical stride effectively resulted in sub-sampling across the vertical direction of the raw input matrix, equivalent to a partial pooling function [144] in addition to the convolution. This resulted in an effective reduction (or overhead) in matrix computing speed that scales inversely with the size of the kernel, so that a 3x3 kernel results in a speed reduction overhead by 1/3. While this can be eliminated by a variety of means to produce convolutions with a symmetric stride and hence no speed overhead, this is actually not necessary for most applications. Finally, this approach is highly flexible and reconfigurable without any change in hardware - we use same system for the convolutional accelerator for image processing as well as to form an optical deep learning CNN which we use to perform a separate series of experiments. The convolutional accelerator hardware forms both the input processing stage as well as the fully connected neuron layer of the CNN (see below). The system can achieve matrix multiplication by simply sampling one-time slot of the output waveform, since the vector dot product is equivalent to the special convolution case where the two input vectors X and W have the same length.

Figure 3 shows a detailed example of the photonic convolution accelerator operating in two different modes. The left panel shows the system performing convolution operations, that are used for the large stand-alone convolution image processing and the convolutional layer of the CNN. The right panel shows the system performing matrix operations which are used as the fully connected layer of the optical CNN. Considering that the experimentally demonstrated configurations are too complex to be presented clearly, in Figure 3 we show a simplified configuration of input data and weights to illustrate the operation principle of our system. The length of W and X shown in this figure are R = 4 and L = 13 for the case of convolution operations, and R = L= 4 for the fully connected layer for matrix operations, respectively.

The schematic of the TOPS photonic convolution accelerator is illustrated in the left panel of Figure 3. The input data vector (length L) and weight vector (length R) is first multiplexed in the time and wavelength domains, respectively. The input data vector is represented by the intensities of the temporal symbols in a stepwise electrical waveform X[n] (n denotes discrete temporal locations of the symbols, $n \in [1, L+R-1]$), where X[n] is the electrical input of the accelerator. The weight vector of the kernel is imprinted onto the optical power of the shaped comb lines as W[R−i+1], at the ith wavelength channel ($i \in [1, R]$, where i increases with wavelength). The input electrical waveform X[n] is first broadcast onto the shaped comb lines via electro-optical modulation. Thus the weighted replica at the ith wavelength channel is W[R−i+1]· X[n]. Next, the optical signals across all wavelengths are progressively shifted in the

time domain via an optical time-of-flight buffer, which provides a wavelength-sensitive (dispersive) delay with a delay step τ (the difference in delay between adjacent wavelengths) equal to the symbol duration (inverse of the Baud rate) of X[n]. Hence, the shifted replica becomes W[R−i+1]· X[n-i]. Finally, the replicas of all wavelengths are summed via photo-detection as

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

where each calculated symbol Y[n] within the range of [R+1, L+1] denotes the dot product between W and a certain region of X (this region is defined by the sliding receptive field as [n−R : n−1] or [n−R, n−R+1, n−R+2, …, n−1]). By simply reading different time slots of the output signal, a convolution is achieved between the weight vector and the input data, thus generating extracted feature maps (matrix convolution outputs) of the input image. While higher order dispersion in the dispersive delay can, in principle, degrade performance, in our experiments this was not a factor.

In addition, the convolution accelerator can also perform matrix multiplication operations, as illustrated in the right panel of Figure 3. The matrix multiplication operations can be treated as a special case of convolution operations when the two input vectors (the pooled and flattened feature maps, and the flattened synaptic weights for the fully connected layer) are the same length (R=L). Figure 3 shows an example with R=L=4. Here, we assume the input data vector XFC[n] and the weight vector WFC[R−i+1] both have a length R (i ∈ [1, R], n ∈ [1, R]). Thus, according to Eq. 1, the output waveform after photodetection is

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

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

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

Considering that the convolutional accelerator fundamentally operates on vectors, for applications to image processing, the input data is in the form of matrices and so it needs to be flattened into vectors. In this work, we follow a common approach where the raw input matrix is first sliced horizontally into multiple sub-matrices, each with a height equal to that of the convolutional kernel. The sub-matrices were then flattened into vectors and connected head-to-tail to form the desired vector. The flattening process for the image processing and the CNN [14] makes the receptive field slide with a horizontal stride of 1 and a vertical stride equal to the height of the convolutional kernel. We note that a small stride (such as the horizontal stride of 1) ensures that all features of the raw data are extracted, while a large stride (eg., a vertical stride of 3 or 5) reduces the overlap between the sliding convolution windows and effectively

subsamples the convolved feature maps, thus partially serving as a pooling function. A stride of 4 was used for the AlexNet [144 – 146].

In addition, we note that although the homogeneous strides are generally used more often in digitally implemented CNNs, inhomogeneous convolution strides (i.e., unequal horizontal and vertical strides) such as those used in this work are also often used and in most cases, such as in our experiments, did not limit the convolution accelerator performance. In our case this was verified by the high recognition success rate of the CNN in full digit prediction. Further, if desired homogeneous convolutions can be achieved by duplicating the weight-and-delay paths (each including a modulator, a spool of dispersive fibre, a de-multiplexer and multiple photo-detectors) of the accelerator. The section below on scaling the network discusses this in more detail.

# 3. Experiment

# 3.1 Optical soliton crystal micro-combs

Optical frequency combs, composed of discrete and equally spaced frequency lines, are extremely powerful for optical frequency metrology [29]. Micro-combs offer the full power of optical frequency combs, but in an integrated form with much smaller footprint [29–35]. They have enabled many breakthroughs in high-resolution optical frequency synthesis [33], ultrahigh-capacity communications [34, 35], complex quantum state generation [36 - 44], advanced microwave signal processing [68 - 88], and more. Figure 4 shows a schematic of our optical microcomb chip as well as typical spectra and pumping curves. We use a class of microcomb called soliton crystals that have a crystal-like profile in the angular domain of tightly packed self-localized pulses within micro-ring resonators [35, 48, 49]. They form naturally in micro-cavities with appropriate mode crossings, without complex dynamic pumping or stabilization schemes (described by the Lugiato-Lefever equation [29, 47]). They are characterized by distinctive optical spectra (Fig. 4f) which arise from spectral interference between the tightly packaged solitons circulating along the ring cavity. Soliton crystals exhibit deterministic generation arising from interference between the mode crossing-induced background wave and the high intra-cavity power (Fig. 4c). In turn this enables simple and reliable initiation via adiabatic pump wavelength sweeping [35] that can be achieved with manual detuning (the intracavity power during pump sweeping is shown in Fig. 4d). The key to the ability to adiabatically sweep the pump is that the intra-cavity power is over 30x higher than single-soliton states (DKS), and very close to that of spatiotemporal chaotic states [29, 35]. Thus, the soliton crystal has much less thermal detuning or instability arising from the ‘soliton step’ that makes resonant pumping of DKS states more challenging. It is this combination of ease of generation and conversion efficiency that makes soliton crystals highly attractive. The coherent soliton crystal microcomb (Figure 4) was generated by optical parametric oscillation in a single integrated MRR (Fig. 4a, 4b) fabricated CMOS-compatible Hydex [23, 24, 35], featuring a Q > 1.5 million, radius 592 μm, and a low FSR of ~ 48.9 GHz. The pump laser (Yenista Tunics – 100S-HP) was boosted by an optical amplifier (Pritel PMFA-37) to initiate the parametric oscillation. The soliton crystal microcomb yielded over 90 channels over the C-band (1540-1570 nm), offering adiabatically generated low-noise frequency comb

lines with a small footprint of < 1 $mm^2$ and low power consumption (>100 mW using the technique in [35]).

# 3.2 Matrix Convolution Accelerator

Figure 2 shows the experimental setup for the full matrix convolutional accelerator to process a classic 500×500 face image. The system performs 10 simultaneous convolutions with ten 3×3 kernels to achieve distinctive image processing functions. The weight matrices for all kernels were flattened into a composite kernel vector W containing all 90 weights (10 kernels with 3x3=9 weights each), which were then encoded onto the optical power of 90 microcomb lines by an optical spectral shaper (Waveshaper), each kernel occupying its own frequency band of 9 wavelengths. The wavelength channels were supplied by a coherent soliton crystal microcomb via optical parametric oscillation in a single micro-ring resonator (MRR) (Fig. 4b), radius 592 μm, FSR spacing ~ 48.9 GHz with an optical bandwidth of ~ 36 nm for 90 wavelengths in the C-band (1540-1570 nm) [35].

Figure 5 shows the experimental image processing results. Figure 5a depicts the kernel weights and the shaped microcomb's optical spectrum while the input electrical waveform of the image (grey lines are theoretical and blue experimental waveforms) are in Figure 5b. Figure 5c displays the convolved results of the 4th kernel that performs a top Sobel image processing function (grey lines are theory and red experimental). Finally, Figure 5d shows the weight matrices of the kernels and corresponding recovered images.

The raw 500×500 input face image was flattened electronically into a vector X and encoded as the intensities of 250,000 temporal symbols with a resolution of 8 bits/symbol (limited by the electronic arbitrary waveform generator (AWG)), to form the electrical input waveform via 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 waveform duration was 3.975μs for each image corresponding to a processing rate for all ten kernels of over 1/3.975μs, equivalent to 0.25 million of these ultra-large-scale images per second.

The input waveform X was then multi-cast onto the 90 shaped comb lines via electro-optical modulation, yielding replicas weighted by the kernel vector W. Following this, the waveform was transmitted through ~2.2 km of standard single mode fibre having a dispersion of ~17ps/nm/km. The fibre length was carefully chosen to induce a relative temporal shift in the weighted replicas with a progressive delay step of 15.9 ps between adjacent wavelengths, exactly matching the duration of each input data symbol τ, resulting in time and wavelength interleaving for all ten kernels.

The 90 wavelengths were then de-multiplexed into 10 sub-bands of 9 wavelengths, each sub-band corresponding to a kernel, and separately detected by 10 high speed photodetectors. The detection process effectively summed the aligned symbols of the replicas (the electrical output waveform of one of the kernels (kernel 4) is shown in Fig. 5c). The 10 electrical waveforms were converted into digital signals via ADCs and resampled so that each time slot of each of the waveforms corresponded to the dot

product between one of the convolutional kernel matrices and the input image within a sliding window (i.e., receptive field). This effectively achieved convolutions between the 10 kernels and the raw input image. The resulting waveforms thus yielded the 10 feature maps (convolutional matrix outputs) containing the extracted hierarchical features of the input image (Figure 5d).

The convolutional vector accelerator made full use of time, wavelength, and spatial multiplexing, where the convolution window effectively slides across the input vector X at a speed equal to the modulation baud-rate — 62.9 Giga Symbols/s. Each output symbol is the result of 9 (the length of each kernel) multiply-and-accumulate operations, thus the core vector computing speed (i.e., throughput) of each kernel is 2×9×62.9 = 1.13 TOPS. For ten kernels computed in parallel the overall computing speed of the vector CA is therefore 1.13×10 =11.3 TOPS, or 11.321×8=90.568 Tb/s (reduced slightly by the optical signal to noise ratio (OSNR)). This speed is over 500 x the fastest ONNs reported to date.

For the image processing matrix application demonstrated here, the convolution window had a vertical sliding stride of 3 (resulting from the 3×3 kernels), and so the effective matrix computing speed was 11.3/3=3.8 TOPs. Homogeneous strides operating at the full vector speed can be readily achieved by duplicating the system with parallel weight-and-delay paths [14], although we found that this was unnecessary. While the length of the input data processed here was 250,000 pixels, the convolution accelerator can process data with an arbitrarily large scale, the only practical limitation being the capability of the external electronics.

To achieve the designed kernel weights, the generated microcomb was shaped in power using two liquid crystal on silicon based spectral shapers (Finisar WaveShaper 4000S). The first flattened the microcomb spectrum while the second located just before the photo-detection performed precise comb power shaping required to imprint the kernel weights. A feedback loop was employed to improve the accuracy of comb shaping, where the error signal was generated by first measuring the impulse response of the system with a Gaussian pulse input and comparing it with the ideal weights.

Figure 6 shows the experimental and theoretical large scale facial image processing results achieved by the matrix convolutional accelerator with ten convolutional kernels. It shows the experimental results of large 500×500 face image processing, including the recorded waveforms and the recovered images. The electrical input data was temporally encoded by an arbitrary waveform generator (Keysight M8195A) and then multicast onto the wavelength channels via a 40 GHz intensity modulator (iXblue). For the 500×500 image processing, we used sample points at a rate of 62.9 Giga samples/s to form the input symbols. We then employed a 2.2 km length of dispersive fibre that provided a progressive delay of 15.9 ps/channel, precisely matched to the input baud rate.

Since there are no common standards in the literature for classifying and quantifying the computing speed and processing power of ONNs, we explicitly outline the performance definitions that we use in characterizing our performance. We follow the approach that is widely used to evaluate electronic micro-processors. The computing power of the convolution accelerator—closely related to the operation bandwidth—is denoted as the throughput, which is the number of operations performed within a certain

period. Considering that in our system the input data and weight vectors originate from different paths and are interleaved in different dimensions (time, wavelength, and space), we use the temporal sequence at the electrical output port to define the throughput in a more straightforward manner.

At the electrical output port, the output waveform has L+R−1 symbols in total (L and R are the lengths of the input data vector and the kernel weight vector, respectively), among which L−R+1 symbols are the convolution results. Further, each output symbol is the calculated outcome of R multiply-and-accumulate operations or 2R OPS, with a symbol duration τ given by that of the input waveform symbols. Thus, considering that L is generally much larger than R in practical convolutional neural networks, the term (L−R+1)/(L+R−1) would not affect the vector computing speed, or throughput, which (in OPS) is given by

$$\frac{2R}{\tau}\cdot\frac{L-R+1}{L+R-1}\approx\frac{2R}{\tau} \qquad (4)$$

As such, the computing speed of the vector convolutional accelerator demonstrated here is 2×9×62.9×10 = 11.321 Tera-OPS for ten parallel convolutional kernels).

We note that when processing data in the form of vectors, such as audio speech, the effective computing speed of the accelerator would be the same as the vector computing speed 2R/ τ. Yet when processing data in the form of matrices, such as for images, we must account for the overhead on the effective computing speed brought about by the matrix-to-vector flattening process. The overhead is directly related to the width of the convolutional kernels, for example, with 3-by-3 kernels, the effective computing speed would be ~1/3 * 2R/τ, which still is in the TOP regime due to the high parallelism brought about by the time-wavelength interleaving technique.

For the convolutional accelerator, the output waveform of each kernel (with a length of L−R+1=250,000−9+1=249,992) contains 166×498=82,668 useful symbols that are sampled out to form the feature map, while the rest of the symbols are discarded. As such, the effective matrix convolution speed for the experimentally performed task is slower than the vector computing speed of the convolution accelerator by the overhead factor of 3, and so the net speed then becomes 11.321×82,668/249,991=11.321×33.07% = 3.7437 TOPS.

For the deep CNN the convolutional accelerator front end layer has a vector computing speed of 2×25×11.9×3 = 1.785 TOPS while the matrix convolution speed for 5x5 kernels is 1.785×6×26/(900−25+1) = 317.9 Giga-OPS. For the fully connected layer of the deep CNN, according to Eq. (4), the output waveform of each neuron would have a length of 2R−1, while the useful (relevant output) symbol would be the one locating at R+1, which is also the result of 2R operations. As such, the computing speed of the fully connected layer would be 2R / (τ*(2R−1)) per neuron. With R =72 during the experiment and ten neurons simultaneous operating, the effective computing speed of the matrix multiplication would be 2R / (τ*(2R−1)) × 10 = 2×72 / (84ps* (2×72−1)) = 119.83 Giga-OPS.

In addition, the intensity resolution (bit-resolution for digital systems) for analog ONNs is mainly limited by the signal-to-noise ratio (SNR). To achieve 8-bit resolution, the SNR of the system needs to be > 20•log10(28) = 48 dB. This was achieved by our accelerator and so our speed in Tb/s is close to the speed in OPs x 8 – not reduced by our OSNR.

# 3.3 Deep Learning Optical Convolutional Neural Network

The convolutional accelerator architecture presented here is fully and dynamically reconfigurable and scalable with the same hardware system. We were thus able to use the accelerator to sequentially form both a frontend convolution processor as well as a fully connected layer, together yielding an optical deep CNN. We applied the CNN to the recognition of full 10 (0-9) handwritten digit images. Figure 7 shows the overall architecture of the 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 shows the architecture of the optical CNN, including a convolutional layer, a pooling layer, and a fully connected layer. Figure 9 shows the detailed experimental schematic of the optical CNN. The left side is the input front end convolutional accelerator while the right is the fully connected layer - both the deep learning optical CNN. The microcomb supplies the wavelengths for both the convolution accelerator as well as the fully connected layer. The electronic digital signal processing (DSP) module used for sampling and pooling is external.

The convolutional layer (Fig. 9, left) performs the heaviest computing duty of the entire network, generally taking 55–90% of the total computing power. The digit images – 30×30 matrices of grey-scale values with 8-bit resolution – were flattened into vectors and multiplexed in the time-domain at 11.9 Giga Baud (time-slot $\tau$ = 84 ps). Three 5×5 kernels were used, requiring 75 microcomb lines, resulting in a vertical stride of 5. The dispersive delay was achieved with ~13 km of SMF to match the data baud-rate. The wavelengths were de-multiplexed into the three kernels which were detected by high speed photodetectors and then sampled and nonlinearly scaled with digital electronics to recover 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 formed the input data to the fully connected layer.

The fully connected layer had 10 neurons, each corresponding to one of the 10 categories of handwritten digits from 0 to 9, with the synaptic weights represented by a 72×10 weight matrix $\mathbf{W}_{FC}^{(l)}$ (ie., ten 72×1 column vectors) for the $l^{th}$ neuron ($l \in [1, 10]$) – with the number of comb lines (72) matching the length of the flattened feature map vector $\mathbf{X}_{FC}$. The shaped optical spectrum at the $l^{th}$ port had an optical power distribution proportional to the weight vector $\mathbf{W}_{FC}^{(l)}$, thus serving as the equivalent optical input of the $l^{th}$ neuron. After being multicast onto the 72 wavelengths and progressively delayed, the optical signal was weighted and demultiplexed with a single Waveshaper into 10 spatial output ports — each corresponding to a neuron. Since this part of the network involved linear processing, the kernel wavelength weighting could be implemented either before the EO modulation or at a later stage just before photodetection. The advantage of the latter is that both the demultiplexing and weighting can then be achieved with a single

Waveshaper. Finally, the different node/neuron outputs were obtained by sampling the 73rd symbol of the convolved results. The final output of the optical CNN was represented by the intensities of the output neurons, where the highest intensity for each tested image corresponded to the predicted category. The peripheral systems, including signal sampling, nonlinear function and pooling, were implemented electronically with digital signal processing hardware, although some of these functions (e.g., pooling) can be performed in the optical domain with the VCA. Supervised network training was performed offline electronically.

We experimentally tested 50 x 8-bit resolution images each 30 × 30 of the handwritten digit dataset with the deep optical CNN. The confusion matrix (Figure 10) shows an accuracy of 88% for the generated predictions, in contrast to 90% for the numerical results calculated on an electrical digital computer. The computing speed of the CA component of the deep optical CNN was 2×75×11.9 =1.785 TOPS, or 14.3 Tb/s. To process image matrices with 5×5 kernels, the convolutional layer had a matrix flattening overhead of 5, yielding an image computing 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 the rate of 1/75.6ns = 13.2 million handwritten digit images per second.

We note that handwritten digit recognition, although widely employed as a benchmark test in digital hardware, for full 10 digit (0 - 9) recognition is still beyond the capability of existing analog reconfigurable ONNs. Digit recognition requires a large number of physical parallel paths for fully-connected networks (e.g., a hidden layer with 10 neurons requires 9000 physical paths), which poses a huge challenge for current nanofabrication techniques. Our CNN represents the first reconfigurable and integrable ONN capable not only of performing high level complex tasks such as full handwritten digit recognition, but at TOP speeds. For the convolutional layer of the CNN, we used 5 sample points at 59.421642 Giga Samples/s to form each single symbol of the input waveform, which also matched with the progressive time delay (84 ps) of the 13km dispersive fibre. The generated electronic waveforms for 50 images [14] served as the electrical input signal for the convolutional and fully connected layers, respectively.

For the convolutional accelerator in both the CA and CNN experiments - the 500×500 image processing experiment and the convolutional layer of the CNN - the second Waveshaper simultaneously shaped and de-multiplexed the wavelength channels into separate spatial ports according to the configuration of the convolutional kernels. As for the fully connected layer, the second Waveshaper simultaneously performed the shaping and power splitting (instead of de-multiplexing) for the ten output neurons. The de-multiplexed or power-split spatial ports were sequentially detected and measured. However, these two functions could readily be achieved in parallel with a commercially available 20-port optical spectral shaper (WaveShaper 16000S, Finisar) and multiple photodetectors. Negative channel weights were achieved using two methods. For the 500×500 image processing experiment and the convolutional layer of the CNN, the wavelength channels of each kernel were separated into two spatial outputs by the WaveShaper according to the signs of the kernel weights, and then detected by a balanced photodetector (Finisar XPDV2020). Conversely, for the fully connected layer the weights were encoded in the symbols of

the input electrical waveform during the electrical digital processing stage. Both of these methods to impart 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 extract the final convolved output. For the CNN, the extracted outputs of the convolution accelerator were further processed digitally, including rescaling to exclude the loss of the photonic link via a reference bit, and then mapped onto a certain range using a nonlinear tanh function. The pooling layer's functions were also implemented digitally, following the algorithm introduced in the network model. The residual discrepancy between experiment and calculations, for both the recognition and convolving functions, was due to the deterioration of the input waveform caused by performance limitations of the electrical arbitrary waveform generator. Addressing this would lead to greater accuracy and closer agreement with numerical calculations.

# 3.4 Network training and digital processing

For the deep learning (multiple level) optical CNN, we employed datasets from the MNIST (Modified National Institute of Standards and Technology) handwritten digit database [144] containing 60000 images as the training set and 10000 images as the test set. The structure of the CNN in this work (Figure 7) was determined empirically using trial-and-error, which is a standard approach for neural networks. In our case this was greatly aided by the fact that the network structure (number of synapses and neurons) could be reconfigured dynamically without any change in hardware. The 28×28 input data was first padded with zeros into a 30×30 image and then sliced into a 5×180 matrix and convolved with the 5×5 kernels. This slicing operation equivalently made the receptive field slide horizontally with a stride = 1 across the rows and a vertical stride = 5 across the columns of the 30×30 input data (corresponding to the 900 input nodes). Then the 6×26×3 feature map was pooled (using average pooling) to a smaller dimension of 6×4×3. Finally, the matrix was further flattened into a 72×1 vector that served as input nodes for the fully connected layer, which in turn generated the predictions using the 10 output neurons. The nonlinear function we used after the convolutional layer, the pooling function and the fully connected layer was the tanh function. Although other nonlinear functions such as ReLU are widely used, we used this tanh function since it can be realized with a saturating electrical amplifier.

The training necessary to acquire pre-trained weights and biases was performed offline with a digital computer. The Back Propagation algorithm [145] was employed to adjust the weights. To validate the hyper-parameters of the CNN, we performed a 10-fold cross validation using the 60000 samples of the training dataset, where the training set was 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 assessed by both the optical CNN (50 images) and an electronic computer (10000 images) for comparison.

# 3.5 Performance comparison

We summarize recent progress of optical neuromorphic hardware in Table 1. This section is not comprehensive but focuses on leading results that address the most crucial technical issues for optical

computing hardware. The input data dimension directly determines the complexity of the processing task. In real-life scenarios, the input data dimension is generally very large, for example, a human face image would require over 60,000 pixels. Thus, to make optical computing hardware eventually useful, the input data dimension would need to be at least over 20,000. In this work we demonstrate processing of images containing 250,000 pixels, which is 224 x higher than previous reports.

The computing speed is perhaps the most important parameter for computing hardware and is the main strength of optical approaches. Although there has not been a widely accepted definition of optical hardware computing speed, the key issue is the number of data sets that are processed within a certain time period - i.e., how many images can be processed per second. As such, although in some approaches [8, 11, 12], the latency is low due to the short physical path lengths, the computing speed remains very low due to the absence of high-speed data interfaces (i.e., input and output nodes are not updated at a high rate). Although other approaches [9, 28] offer high-speed data interfaces, their computing parallelism is not high and so their speed is similar to the input data rate. In our work, [14] through the use of high-speed data interfaces (62.9 Giga Baud) and time-wavelength interleaving, we achieved a record computing speed of 11.321 Tera-OPS, > 500 x higher than previous reports.

Finally, the scalability and reconfigurability determines the versatility of the optical computing hardware. Approaches that cannot dynamically reconfigure the synapses [11] (marked as “Level 1” in the table) are barely trainable. Approaches at Level 2 [9, 12, 28] support online training, however, they can only process a specific task since the network structure is fixed once the device is fabricated. For approaches [28] at Level 3, different tasks can be processed although the function of each layer is fixed, which limits the hardware from implementing more complex operations other than matrix multiplication. Our work represents the first approach that operates at Level 4 with full dynamic reconfigurability in all respects. Here, the synaptic weights can be reconfigured by programming the WaveShaper. Further, the number of synapses per neuron can be reconfigured by reallocating the wavelength channels with the de-multiplexer. The number of layers can be reconfigured by changing the number of stacked devices. Finally, the computing function can be switched between convolution and matrix multiplication by changing the sampling method. The degree of integration directly determines the potential computing density (processing capability per unit footprint). For approaches not well suited to integration [8, 11, 28], the potential computing density is low. While other approaches achieve limited integration of the weight and sum circuits [8, 12] - probably the most challenging issue — advanced integrated light sources have not been demonstrated. The performance of the light source directly determines the performance of the overall hardware in both input data scale [8] and number of synaptic connections per neuron [12]. The $mm^2$ sized microcomb offers a large number of precisely-spaced wavelengths, which enhances the overall parallelism and computing density, representing a major step towards the full integration of optical computing hardware.

# 4. Scaling The Network

This approach can be readily scaled in performance in terms of input data size, as well as network size and speed. The data size is limited in practice only by the memory of the electrical digital-to-analog converters, and so in principle it is possible to process 4K-resolution (4096×2160) images. By integrating 100 photonic convolution accelerators layers (still much less than the 65536 processors integrated in the Google TPU [22]), the optical CNN would be capable of solving much more difficult image recognition tasks at a vector computing speed of 100 × 11.3=1.130 Peta-OPS. Further, the optical CNN presented here supports online training, since the optical spectral shaper used to establish the synapses can be dynamically reconfigured as fast as 500 ms or faster with integrated optical spectral shapers [147].

Although we had a non-trivial optical latency of 0.11 μs introduced by dispersive fibre spool, this did not affect the operational speed. Moreover, the latency of the delay function can be virtually eliminated (to < 200 ps) by using integrated highly dispersive devices such as photonic crystals or customized chirped Bragg gratings [148] or even tunable dispersion compensators [149, 150]. Finally, current nanofabrication techniques can enable significantly higher levels of integration of the convolutional accelerator. The micro-comb source itself is based on a CMOS compatible platform that is intrinsically designed for large-scale integration. Other components such as the optical spectral shaper, modulator, dispersive media, de-multiplexer and photodetector have all been realized in integrated form [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 and equipment. 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 – only subject to tradeoffs in signal-to-noise ratio (SNR). 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.

# 5. Conclusion

We demonstrate a universal optical vector convolutional accelerator operating beyond 10 TOPS, generating convolutions of images of 250,000 pixels with 8-bit resolution for 10 kernels simultaneously — enough for facial image recognition. We then use the same hardware to sequentially form a deep optical CNN with ten output neurons, achieving successful recognition of full 10 digits with 900 pixel handwritten digit images with 88% accuracy. Our approach is scalable and trainable to more complex networks for demanding applications to unmanned vehicles and real-time video recognition.

# Declarations

Competing interests: The authors declare no competing interests.

# 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 <background-color:#FFCC66;bu>37</background-color:#FFCC66;bu> (2) 338–347 (2019).
91. S. Sciara et al., “Generation and Processing of Complex Photon States with Quantum Frequency Combs”, IEEE Photonics Technology Letters <background-color:#FFCC66;bu>31</background-color:#FFCC66;bu> (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. <background-color:#FFCC66;bu>42</background-color:#FFCC66;bu> (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**, **K**⍰**stutis 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**)**.

# Table 1

Table 1 can be found in the supplementary files section.

# Figures

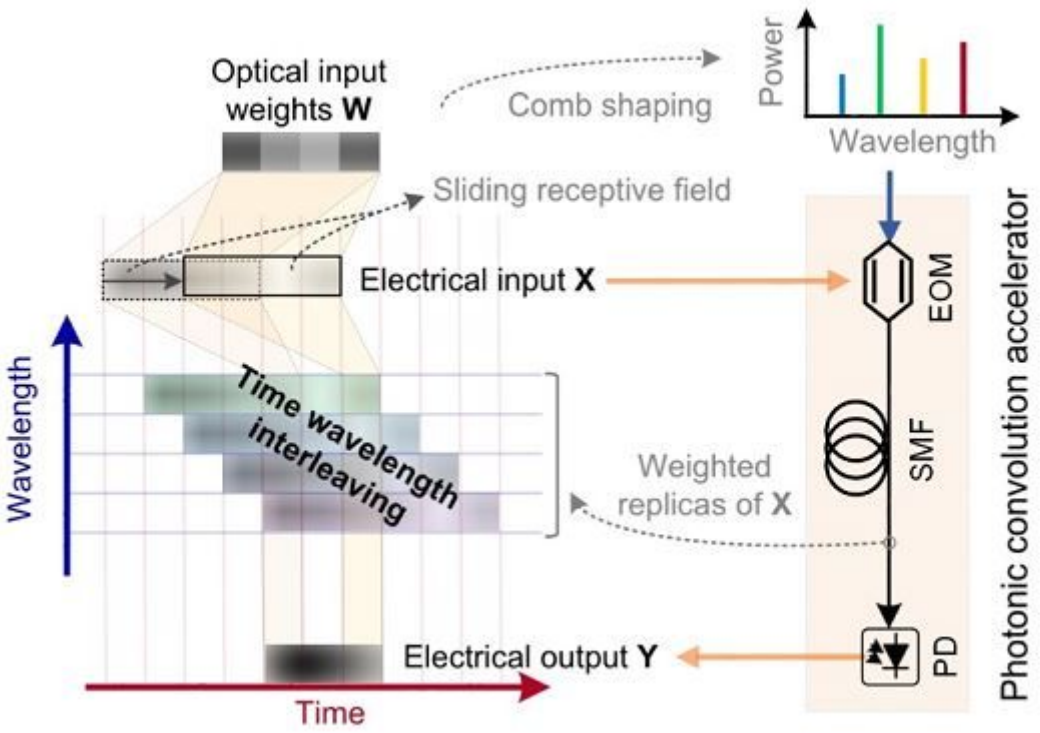
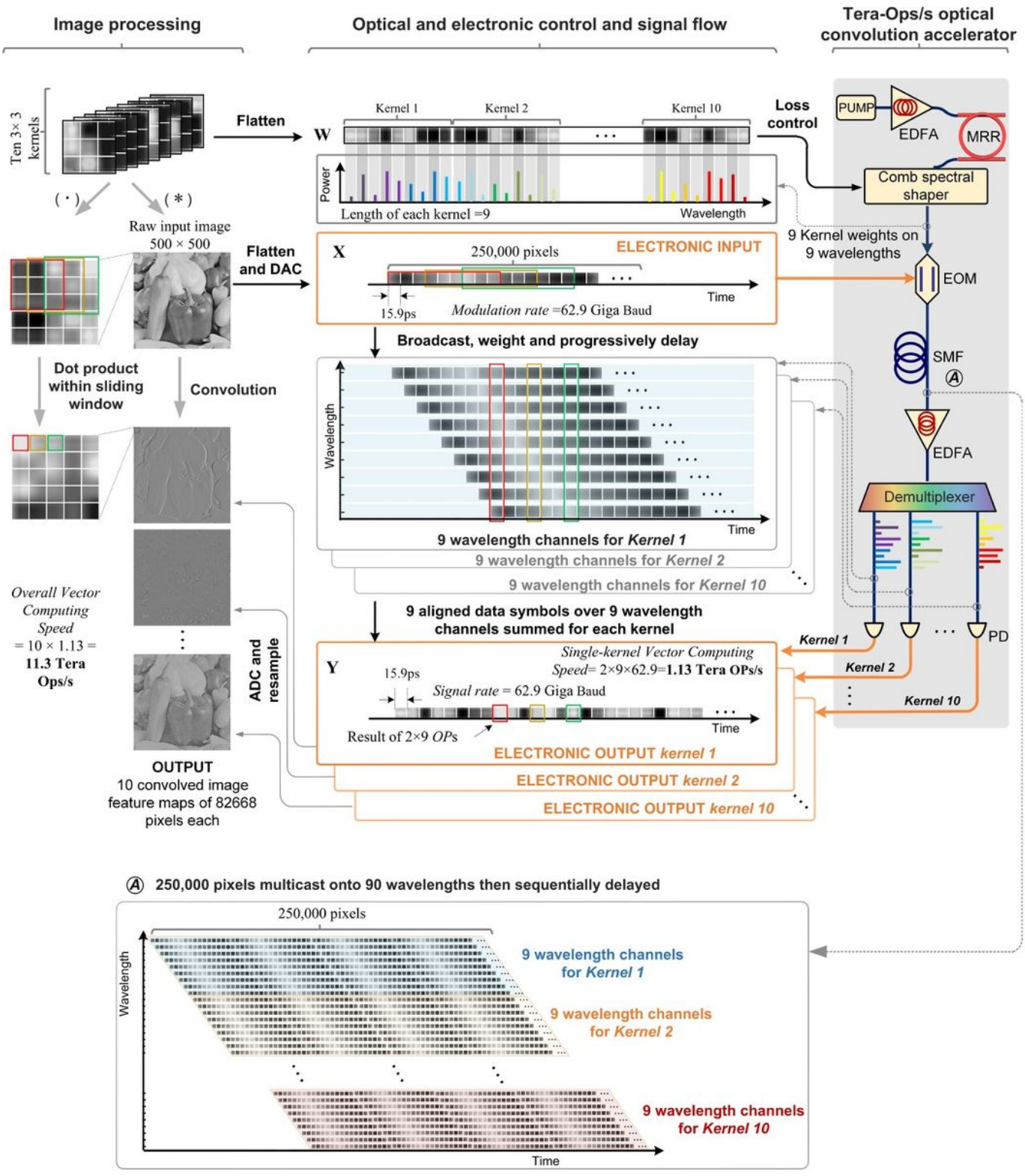


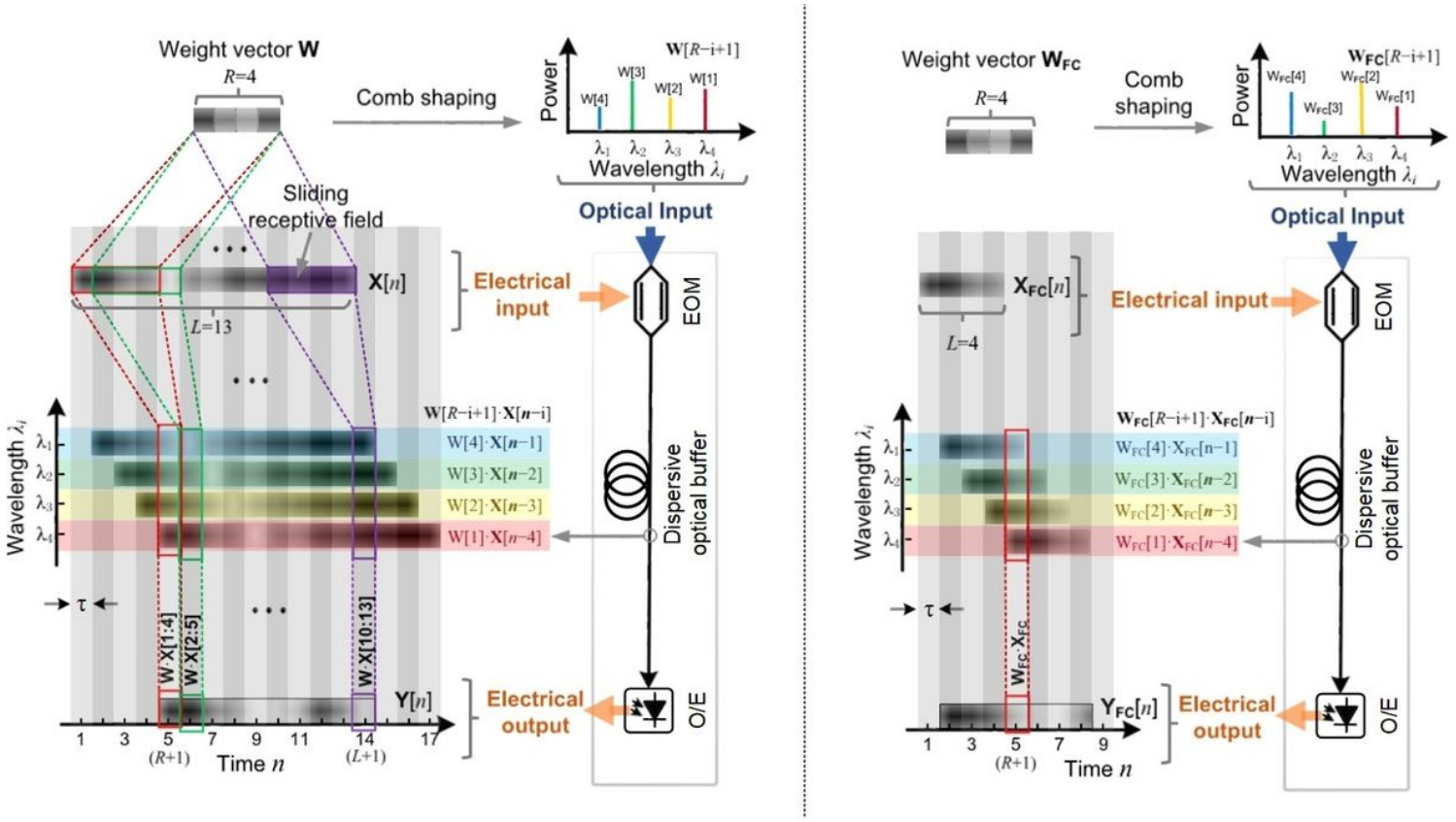

**Figure 1**

**Operation principle of the Tera-FLOPS photonic convolution accelerator**. EOM: electro-optical Mach-Zehnder modulator. SMF: standard single mode fibre for telecommunications. PD: photodetector.



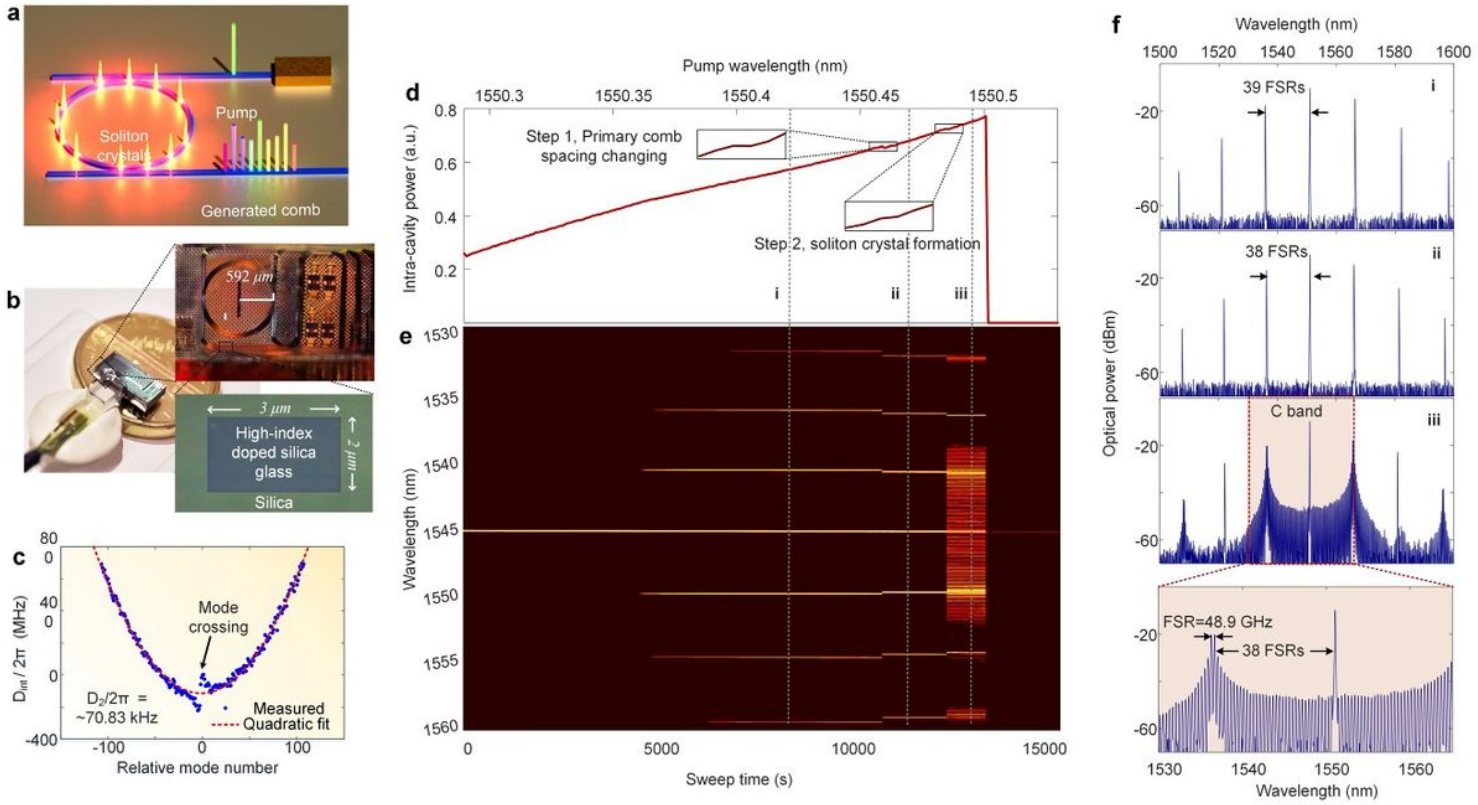

## Figure 2

**Image processing**, consisting of the experimental setup (right panel), the optical and electronic control and signal flow (middle panel), and the corresponding 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.



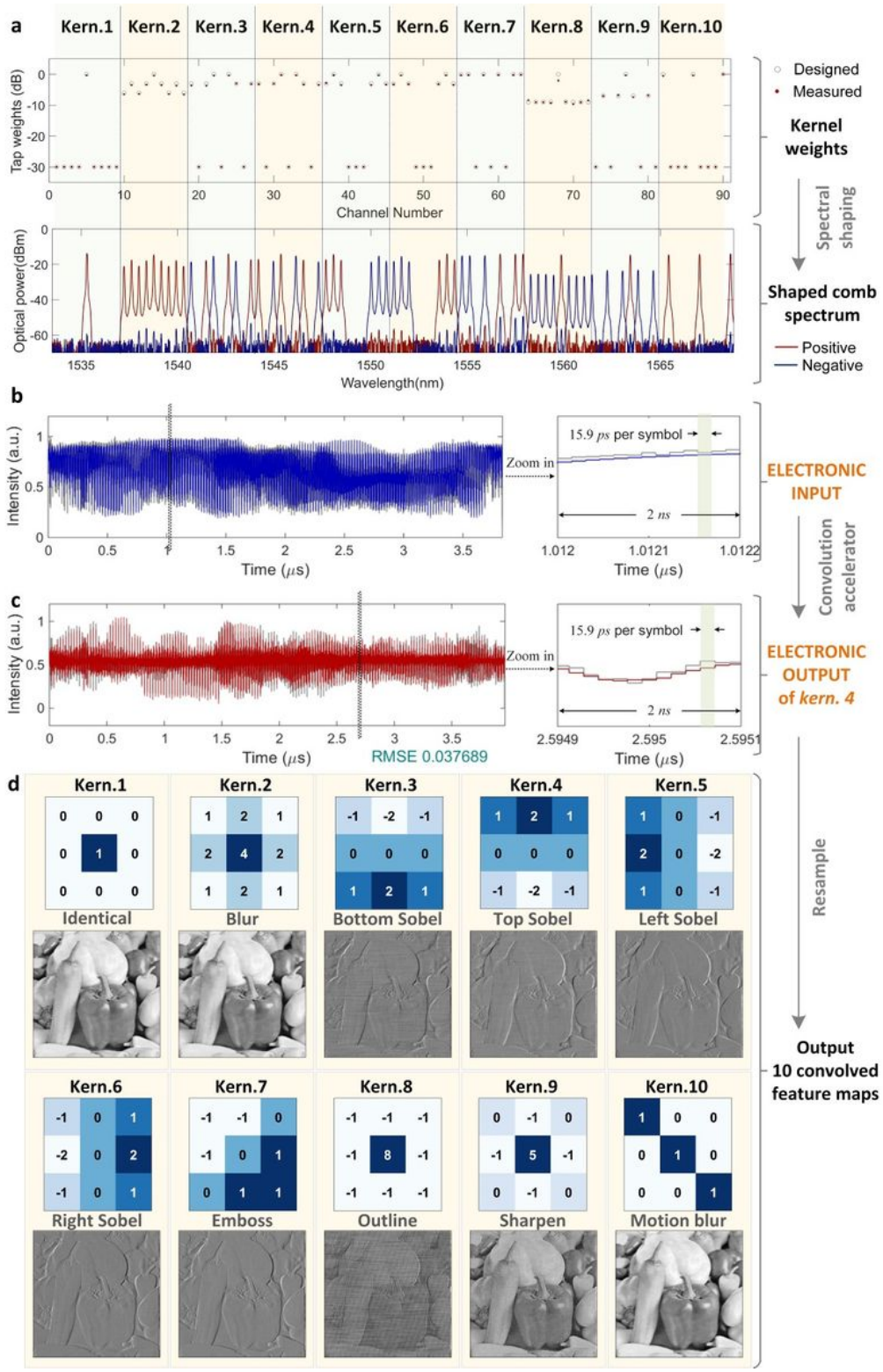

## Figure 3

**TOPS photonic convolution accelerator operating in two modes.** Left: convolution operation mode with $R$ = 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).



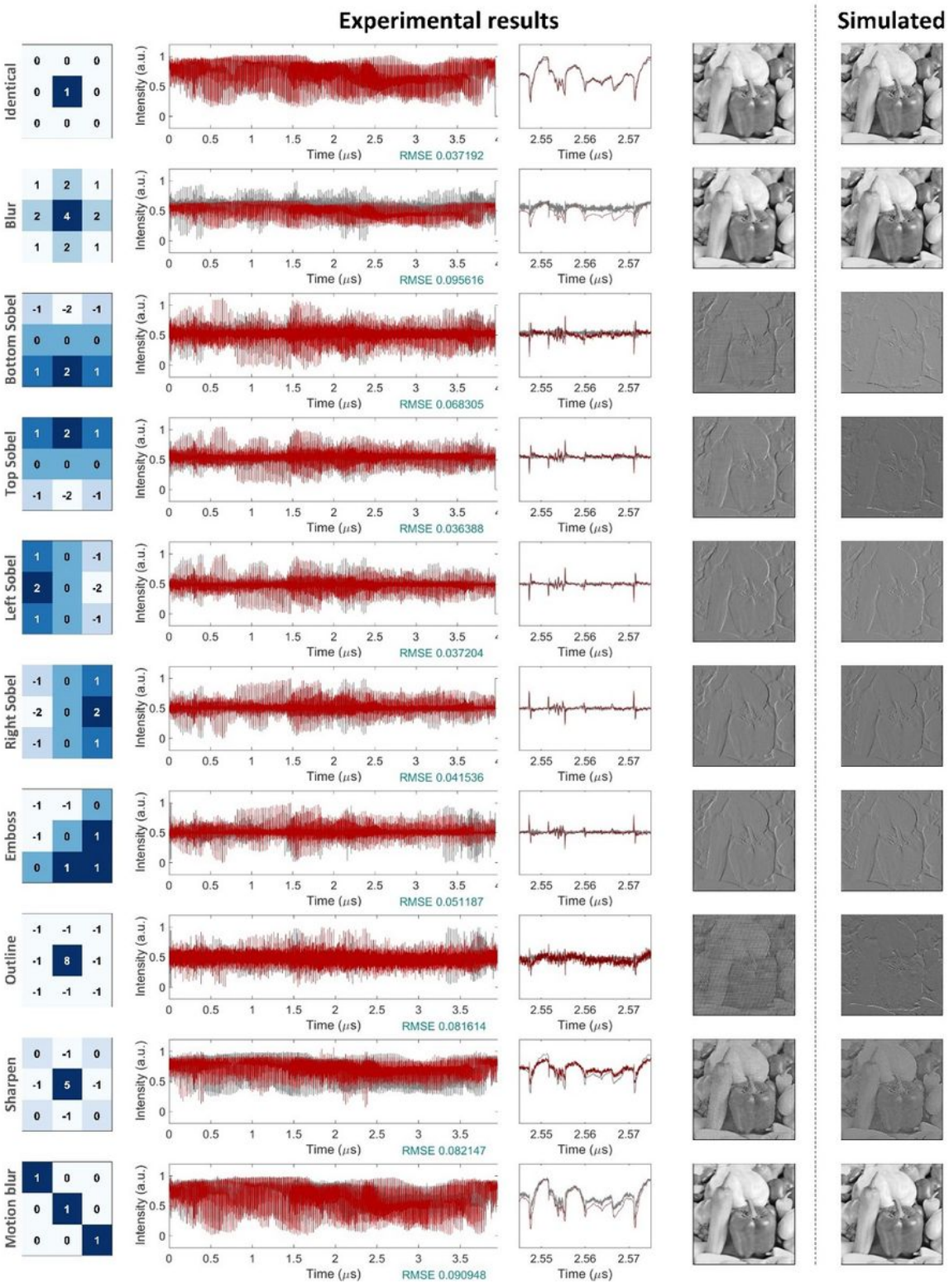

## Figure 4

**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.




## 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 4$^{th}$ 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 achieved by the convolutional accelerator with ten convolutional kernels.

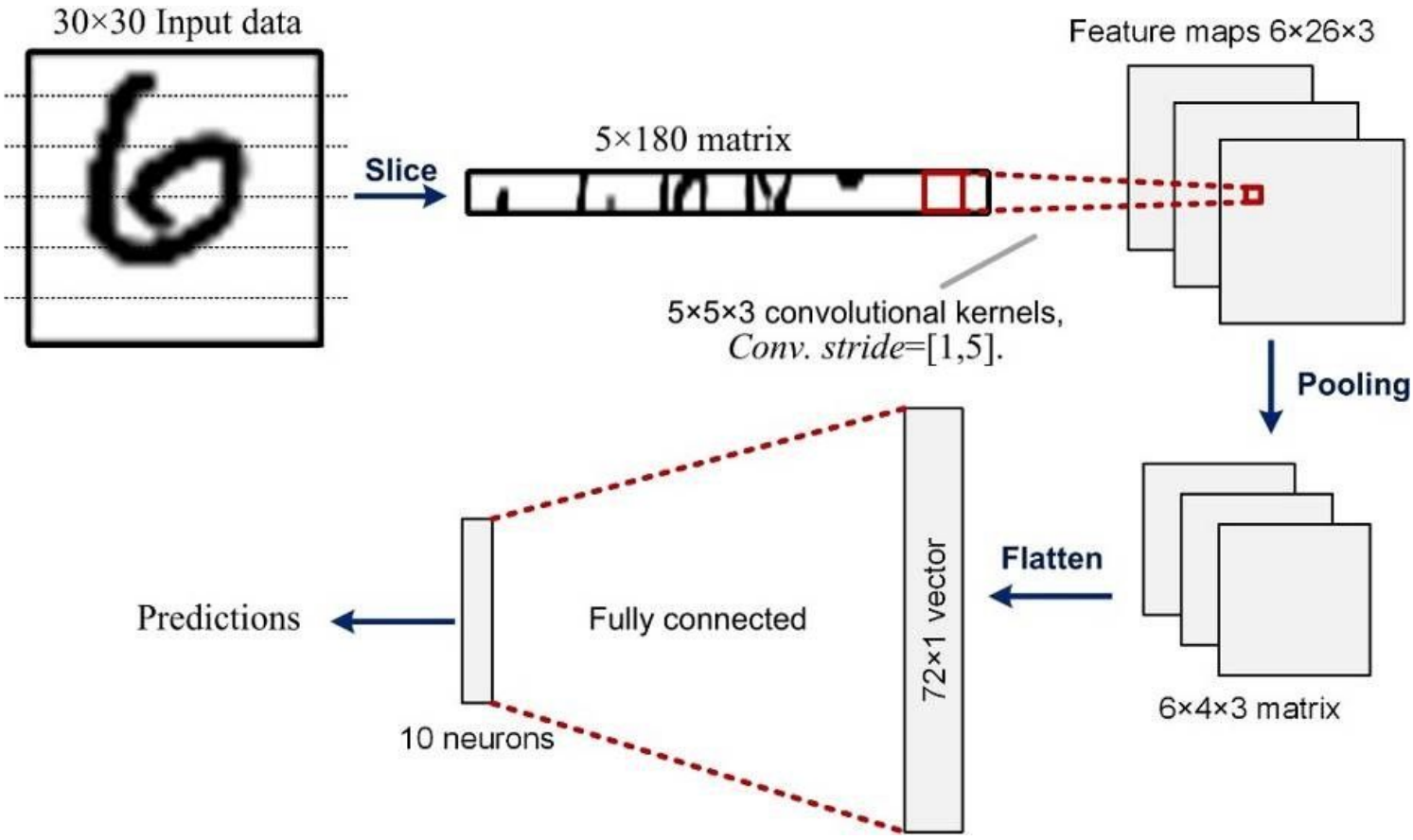
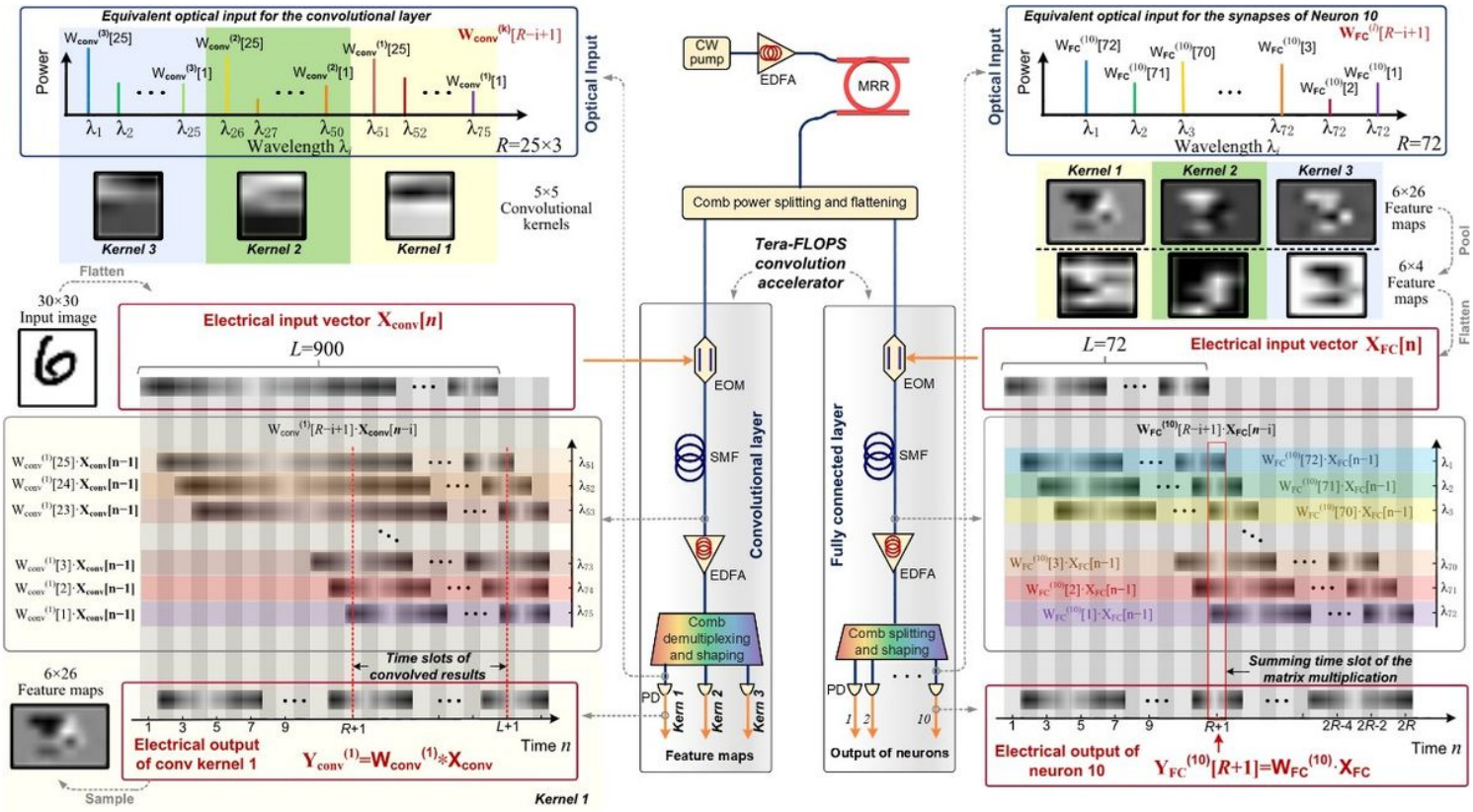



**Figure 7**

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 a convolutional layer, a pooling layer, and a fully connected layer.




## Figure 9

**Experimental schematic of the optical CNN.** Left side is the input front end convolutional accelerator while the right side is the fully connected layer, both of which form the deep learning optical CNN. The microcomb source supplies the wavelengths for both the tera-OPS photonic convolution accelerator as well as the fully connected layer systems. The electronic digital signal processing (DSP) module used for sampling and pooling etc. is external to this structure.

## Figure 10

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

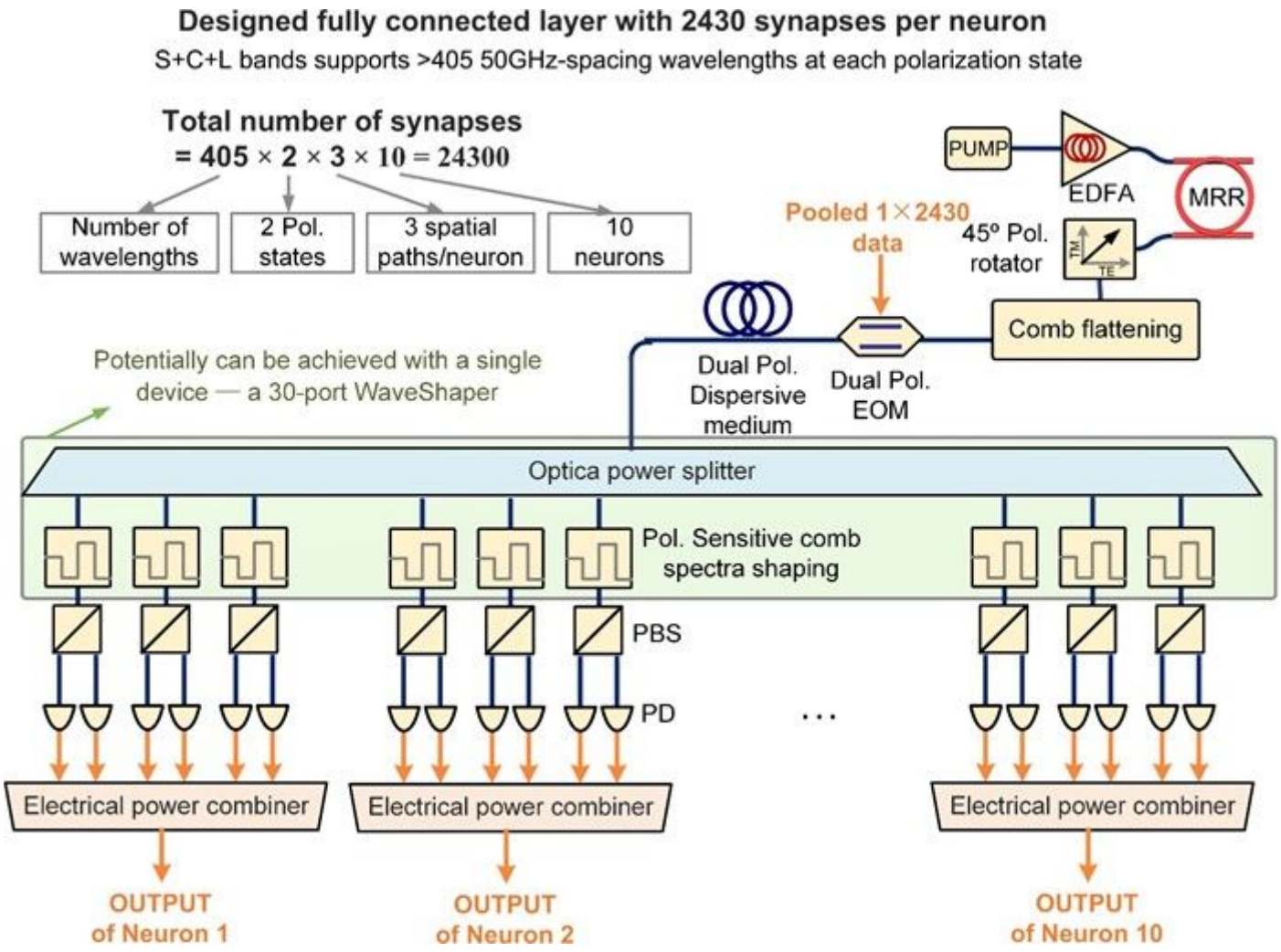



**Figure 11**

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.

# Supplementary Files

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

- table1.jpg