Supplementary information for

Photonic embedding learning with high energy efficiency exceeding 100 GOPS/W/mm^2

Yuyao Huang[1,2,†], Wencan Liu[1,2,†], Run Sun[1,2], Peng Meng Chan[3], Yutong He[1,2], Tianle Chen[1,2], Sigang Yang[1,2], Tingzhao Fu[4,*], and Hongwei Chen[1,2,*]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

[2]Beijing National Research Center for Information Science and Technology (BNRist),

[3]School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China

[4]Hunan Provincial Key Laboratory of Novel Nano Optoelectronic Information Materials and Devices, College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China

[†]These authors contribute equally to this work.

[*]Correspondence: chenhw@tsinghua.edu.cn, futingzhao@nudt.edu.cn

1 Supplementary note 1: Optimization of the diffractive core

In our previous work, we designed the diffractive core with a propagation length of 300 um between metalines. This design ensures a constant effective refractive index for the silica slot group, regardless of the incident angle of incoming light, enabling consistent phase modulation functionality across metalines at different positions. The rationale behind this extended propagation length lies in allowing the spherical wave emitted from each input waveguide to evolve into a wavefront analogous to a plane wave after sufficient propagation. This transformation ensures that the phase delay induced by the metalines remains uniform across all positions of the incident wavefront. However, this long distance between metalines layers also brings high insertion loss. In [1], the optical loss of dffractive core with 2 layers of metalines with propagation length of 300 um is 15.11 dB.

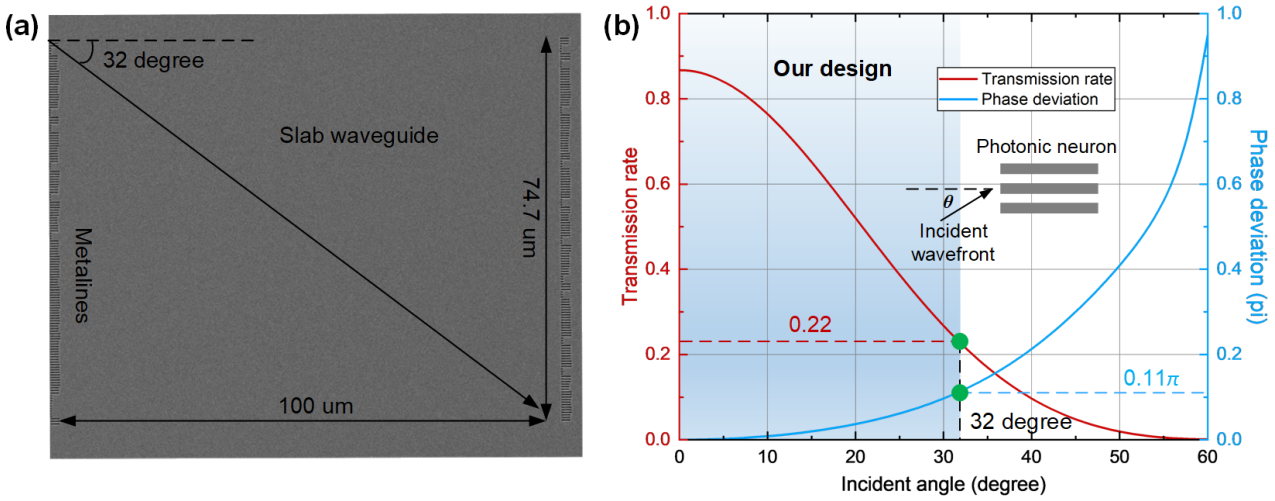

Figure S1: (a) Maximum incident angle for metalines with a diffractive propagation length of 100 um. (b) Transmission and phase deviations as a function of varying incident light angles to the photonic neuron.

Minimizing the propagation distance can significantly reduce insertion loss. However, shortening this distance may introduce phase and transmission errors in the pre-trained metalines, as variations in the incident light's angle across different metaline positions prevent the wavefront from approximating a plane wave. To address this issue, we investigate the phase and transmission responses of the silica slot for incident light at varying angles, as illustrated in Fig.S1. As the incident angle varies from 0 to 60 degrees, the transmission rate of each silica slot decreases, while additional phase noise increases, as shown in Fig.S1(b). For our design with a 100 um propagation length between metalines as shown in Fig.S1(a), the transmission rate drops to as low as 0.22, introducing a phase deviation of up to 0.11π. To characterize the impact of the incident angle on transmission and phase delay, we polynomially fit their respective curves as described in Eq.S1 and Eq.S2. By applying a complementary term, $Tr(\theta) \cdot \exp(\mathrm{j} \cdot \Delta\psi(\theta))$, into the response of photonic neurons in metalines for incident light with an angle of θ, the errors introduced by reducing the footprint of the diffractive core can be mitigated during the modeling and training process. Using this optimized modeling, we fabricate PEUs with a more compact footprint, achieving a measured optical loss of 9.52 dB, representing an improvement of approximately 6 dB compared to our previous work. This insertion loss can be further reduced by incorporating the fitting functions for transmission and phase noise as a function of incident angle.

$$Tr(\theta) = -3 \times 10^{-7} \cdot \theta^4 + 4 \times 10^{-5} \cdot \theta^3 - 0.0017 \cdot \theta^2 + 0.0039 \cdot \theta + 0.8622 \tag{S1}$$

$$\psi(\theta) = 2 \times 10^{-7} \cdot \theta^4 - 2 \times 10^{-5} \cdot \theta^3 + 0.0008 \cdot \theta^2 - 0.0077 \cdot \theta + 0.0176 \tag{S2}$$

2 Supplementary note 2: Experimental results for PEU in grayscale image compression

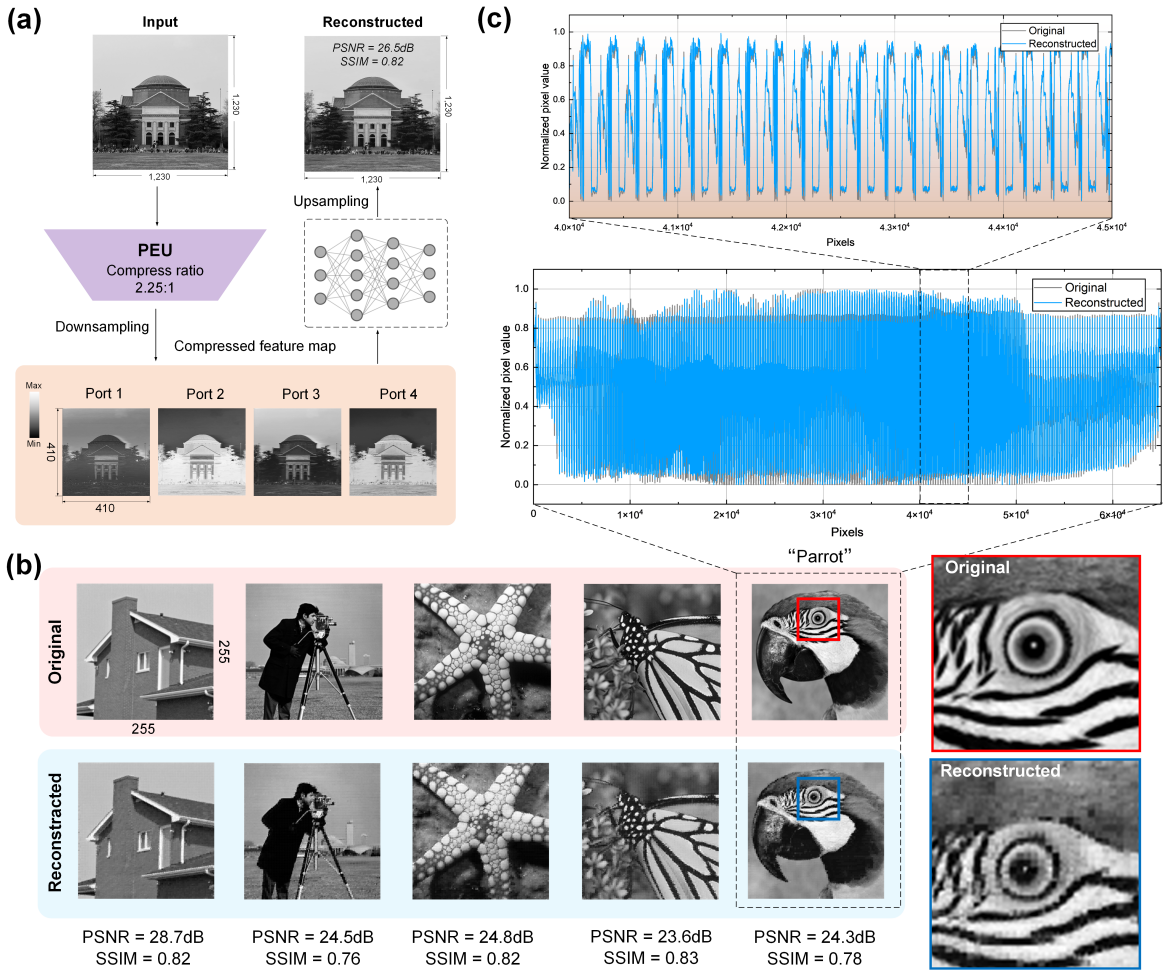

Figure S2: (a) Expeimental pipline of Megapixel grayscale image compression and reconstruction. (b) Comparison of the original and compressed-reconstructed images from the SET12 dataset, along with the corresponding PSNR and SSIM. (c) Pixel-by-pixel comparison between the original and reconstructed image.

To validate the proposed PEU's capabilities, we fabricated a PEU chip for a grayscale image compression task [2]. Unlike color images, grayscale images contain only luminance information, simplifying the processing requirements. In our experimental demonstration, image patches are modulated using thermo-optic phase shifters exclusively, compressing each 3×3 patch into 2×2, achieving a compression ratio of 2.25:1. The compressed data is reconstructed using a lightweight electrical upsampling network,

as illustrated in Fig.S3(a). Fig.S3(b) presents the experimental results of the PEU's performance on grayscale image compression using the SET12 dataset, with the corresponding PSNR and SSIM values provided. Additionally, Fig.S3(c) shows a comparison of the serialized pixel values for the "Parrot" image from the dataset and its reconstructed version.

3 Supplementary note 3: Experimental feature maps for CSET8 image compression

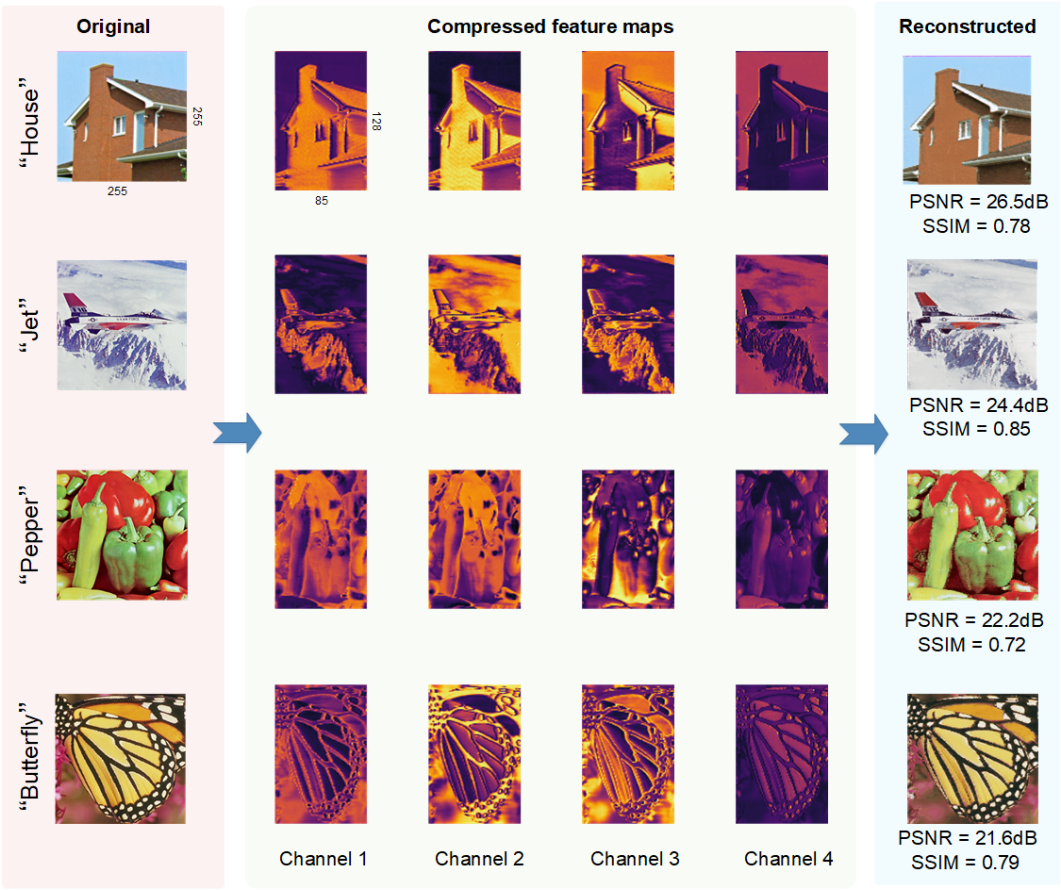

Figure S3: Experimental results of the PEU-based image compression and reconstruction system on color images from the CSET8 dataset

4 Supplementary note 4: Photonic channel-wise pooling

In deep convolutional neural networks (CNNs) such as VGG, AlexNet, and ResNet, multi-kernel convolution layers and pooling layers are widely employed to reduce the dimensions of individual feature maps. However, as the network depth increases, the number of feature map channels often grows dramatically, reaching 512, 1024, or even higher. This exponential growth poses significant computational challenges for hardware-based back-end processing. To overcome this bottleneck, channel-wise pooling [3] has been introduced as an effective dimensionality reduction technique. It reduces the number of channels by applying multiple 1×1 kernels between convolutional layers, where feature maps are convolved across

channels. This approach decreases the number of features, compresses data, and reduces the parameter size of subsequent network layers, as illustrated in Fig.S4(a).

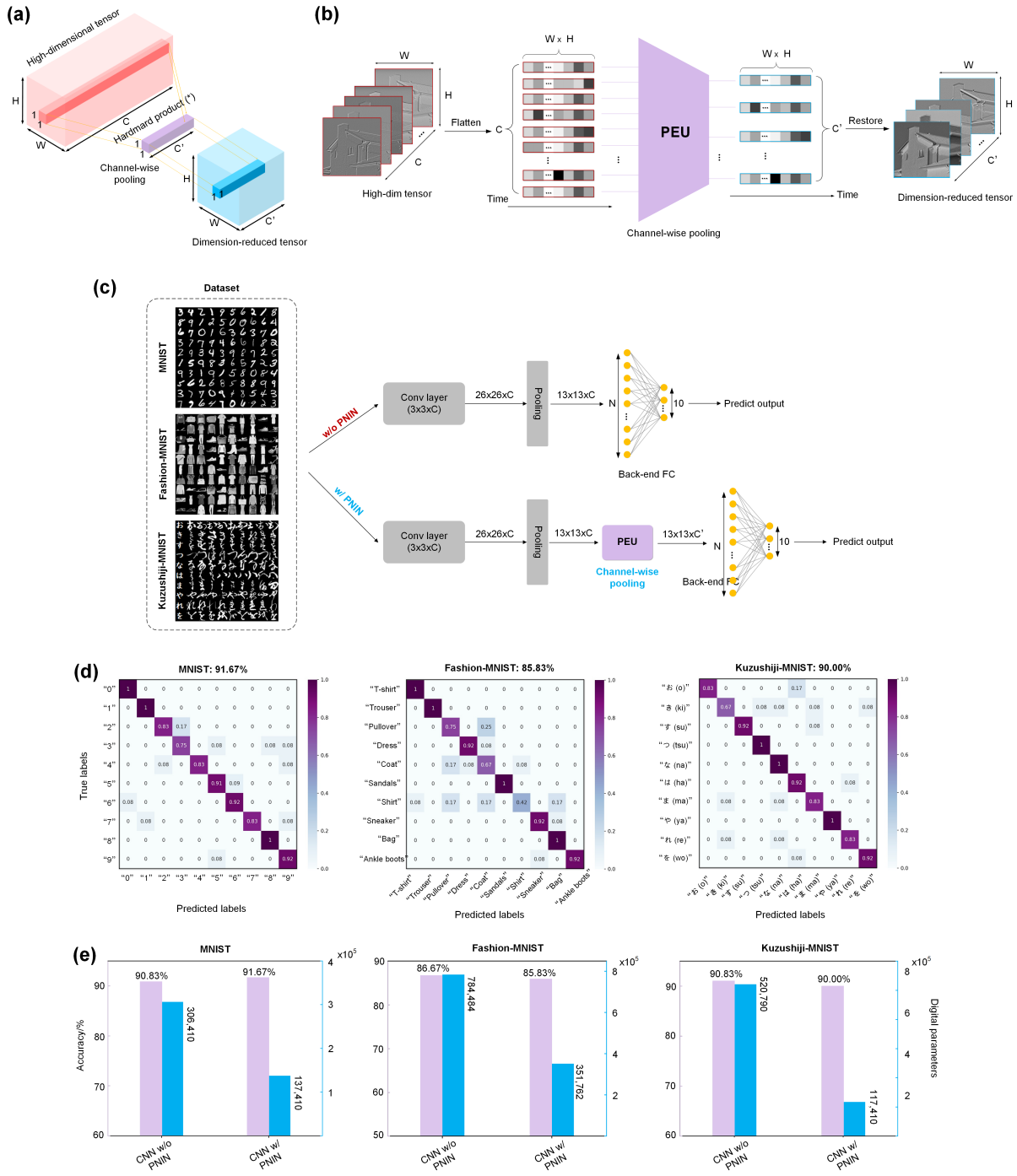

Figure S4: (a) Principle of channel-wise pooling. (b) Photonic channel-wise pooling architecture based on PEU. (c) Experimental pipline for validating the photonic channel-wise pooling. (d) The confusion matrices of the three datasets in classification with PEU-based photonic channel-wise pooling. (e) The test accuracy and the corresponding fully connected network parameter size when the CNN is tested with and without photonic channel-wise pooling.

We demonstrate the application of the PEU for photonic channel-wise pooling, as illustrated in Fig.S4(b). High-dimensional tensors of size $W \times H$ with channel size C are flattened into C vectors and loaded onto the PEU chip via complex-valued modulation. At the output, C' waveguides ($C' < C$) enable pixels from different tensor channels to diffract and interfere, effectively compressing the tensor dimensions while preserving essential information. For the experiment, we selected three datasets: MNIST, Fashion-MNIST, and Kuzushiji-MNIST, as shown in Fig.S4(c). The image datasets were processed using two CNNs for classification, each consisting of a convolutional layer ($3 \times 3 \times C$), a pooling layer (3×3), and a back-end fully connected network. To evaluate photonic channel-wise pooling, we integrated the PEU chip into one of the networks. The PEU was configured with C inputs and C' outputs to perform channel-wise pooling and feature map fusion. For MNIST and Fashion-MNIST, the PEU was configured with $C = 9$ using phase shifters to load the data. For Kuzushiji-MNIST, $C = 18$ was used, employing complex-valued modulation. The confusion matrix results for the three datasets when PEU is applied are presented in Fig.S4(d). Fig.S4(e) shows the test accuracy and the corresponding parameter sizes of the fully connected network with and without PEU. Notably, the parameter size of the back-end fully connected network is reduced by 56%, 56%, and 78% for MNIST, Fashion-MNIST, and Kuzushiji-MNIST, respectively, while maintaining accuracy comparable to that of a purely electronic CNN without PEU.

5 Supplementary note 5: Computational density of PEU

The fabricated PEU chip consists of three fully connected layers, each containing N photonic neurons, with I input ports and O output ports. The total number of operations (OP) includes the connections between the input-to-hidden layers, hidden-to-hidden layers, and hidden-to-output layers. Consequently, the number of multiplications equals to the total number of connections, while the number of additions equals to the total number of connections minus the number of connected nodes. Furthermore, within the hidden layers, each neuron contributes a complex-valued coefficient. Operating at a specific modulation speed f and comprising three layers with N designed photonic neurons per layer, the computational throughput (CT) of the PEU can be calculated as shown in Eq. (1).

$$CT = f \times [(2I - 1) \times N + (2N - 1) \times N + (2N - 1) \times O + 2N] OPS \tag{1}$$

In our experimental setup in Section 2.4, the fabricated PEU includes 9 input ports, 4 output ports, and each layer contains 50 photonic neurons. The experimental setup uses thermo-optic modulation with frequency of 20 kHz. Thus, CT of PEU chip can be calculated as 125.92 MOPS. The computational core of PEU chip has a dimension of 300 um in total length and 74.7 um in width, resulting in an footprint of 0.02 mm^2. Consequently, the computational capacity of the PEU chip is $125.92/0.02 = 6.296$ GOPS/mm^2. The primary limitation lies in the relatively low modulation rate of the thermal optical modulation scheme employed. Assuming the integration of plasma dispersion effect-based silicon modulators in our design with a moderate 10 GHz modulation speed, the calculated density can be increased to 3.148 POPS/mm^2.

6 Supplementary note 6: Computational energy efficiency of PEU

The power consumption of PEU chip mainly comes from five parts: external laser source, on-chip complex-valued modulator array, external current source, photodetectors, temperature electrical controller (TEC) and digital backend. We conclude the power consumption of the PEU chip in Table S1.

Fig. S5(a) illustrates the cross-sectional structure of the fabricated PEU chips. The average resistance of the TiN heaters is measured at 178.4 Ω. For the on-chip thermal phase shifters, the average tuning current is 3 mA, while achieving a π-phase shift in the TOMZIs requires 9.2 mA. Quadrature points are fixed at 5.4 mA, as shown in Fig. S5(b). This results in a total power consumption of the complex-valued modulators of 0.061 W. Additionally, four photodetectors are utilized, each with a power consumption of 0.5 W, contributing a total of 2 W. A TEC submount packaged beneath the PEU chip operates at 0.15 V and 0.08 A, consuming 0.012 W. The digital processor, which handles control, feature mapping, and back-end network operations, consumes an estimated 30 W, powered by an Intel Core i7-11800H CPU. An external laser source consumes 50 W. Together, the total system power consumption is approximately 82.07 W.

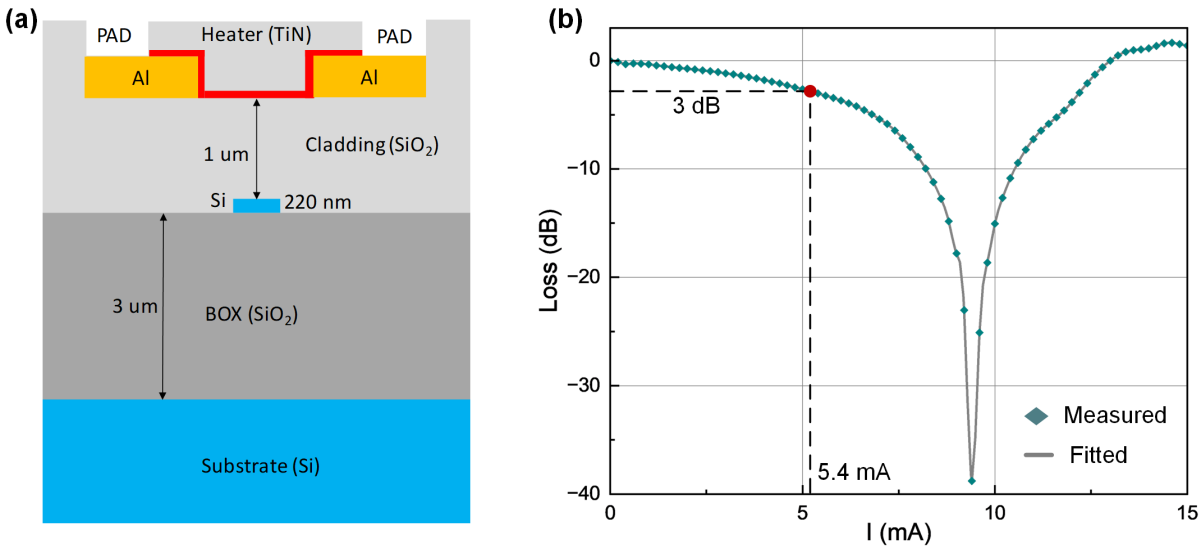

Figure S5: (a) Cross-sectional structure of the fabricated PEU chips. (b) Transmission curve of the TOMZIs under varying applied current.

The total energy consumption per operation (ECPO) of the system for a single inference can therefore be approximated as Eq. (2):

$$ECPO = \frac{Power\ consumption}{MS \times Inference\ operations} \tag{2}$$

where $Power\ consumption$ is the total power of the photonics, drivers, and benchtop electronics and MS is the modulation speed, corresponding to a single inference. $Inference\ operations$ is the total operations in a single inference. Our devices perform 6,296 operations per inference as calculated above, resulting a total on-chip ECPO of 0.061 W/(20 kHz × 6,296 OPs) = 484.43 pJ/OP and 484.43 pJ/OP × 0.02mm^2 = 9.69 pJ·mm^2/OP. The total ECPO including all power devices can be calculated as 82.07 W/(20 kHz × 6296 OPs) = 0.65 μJ/OP. That is, the on-chip energy efficiency of the PEU can be calculated as

103.21 GOPS/(W·mm^2). The overall energy efficiency which including benchtop instruments is 76.71 MOPS/(W·mm^2).

Assuming the integration of carrier dispersion effect-based silicon modulators operating at a moderate 10 GHz modulation frequency with an energy efficiency of 10 fJ/bit, and four on-chip Ge photodetectors with a responsivity of 1.12 A/W and receiving power of -20 dBm consuming 0.045 W, the on-chip power efficiency can be significantly enhanced. Discrete packaged DACs, such as the Analog Devices AD8802 [4], can further reduce power consumption for phase shifters encoding the weights, consuming approximately 5 uW per channel. Additionally, a simple digital back-end network implemented on advanced electrical analog chips consumes less than 1 µW [5]. Moreover, a packaged tunable telecom laser operating at 20 mW with 20% wall-plug efficiency adds only 0.12 W to the power budget. With these optimizations, the total power consumption is estimated at 1.39 W. This achieves a calculated ECPO of 22 fJ/OP and an overall power efficiency of 2.26 POPS/(W·mm^2).

Table S1: **Estimated power consumption of the experimental system of PEU.**

Components	Power (W)	Estimated Power (W)
External laser source	50 (benchtop)	0.12[6]
Digital controller	30 (benchtop)	0.09[4]
Digital backend network (powered by intel core i7-11800H)		0.001[5]
On-chip complex-valued modulator array	0.061	1.13*
TEC	0.012	
Photodetectors	2	0.045**
Total power consumption	82.07	1.39

*10 fJ/b energy efficiency for modulators with 10 GHz frequency is taken for the estimation[7].

**1.12 A/W Ge photodetector with recieving power of -20 dBm is taken for the estimation[8].

References

[1] Yuyao Huang, Wencan Liu, Run Sun, Tingzhao Fu, Yaode Wang, Zheng Huang, Sigang Yang, and Hongwei Chen. Diffraction-driven parallel convolution processing with integrated photonics. *Laser & Photonics Reviews*, page 2400972, 2024.

[2] Xiao Wang, Brandon Redding, Nicholas Karl, Christopher Long, Zheyuan Zhu, James Skowronek, Shuo Pang, David Brady, and Raktim Sarma. Integrated photonic encoder for low power and high-speed image processing. *Nature Communications*, 15(1):4510, 2024.

[3] M Lin. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[4] Ad8802 datasheet, 2022. https://www.analog.com/media/en/technical-documentation/data-sheets/AD8802_8804.pdf.

[5] Yitong Chen, Maimaiti Nazhamaiti, Han Xu, Yao Meng, Tiankuang Zhou, Guangpu Li, Jingtao Fan, Qi Wei, Jiamin Wu, Fei Qiao, et al. All-analog photoelectronic chip for high-speed vision tasks. *Nature*, 623(7985):48–57, 2023.

[6] Saumil Bandyopadhyay, Alexander Sludds, Stefan Krastanov, Ryan Hamerly, Nicholas Harris, Darius Bunandar, Matthew Streshinsky, Michael Hochberg, and Dirk Englund. Single-chip photonic deep neural network with forward-only training. *Nature Photonics*, 18(12):1335–1343, 2024.

[7] Guoliang Li, Xuezhe Zheng, Jin Yao, Hiren Thacker, Ivan Shubin, Ying Luo, Kannan Raj, John E Cunningham, and Ashok V Krishnamoorthy. 25gb/s 1v-driving cmos ring modulator with integrated thermal tuning. *Optics Express*, 19(21):20435–20443, 2011.

[8] Amf-qp-rnd-011 amf pdk3.1 user manual v1, 2020. http://www.advmf.com/.