Problem: Optical neurons (e.g., in D2NNs) must be physically fabricated with quantized, bounded phase values, causing model mismatch and major performance loss when full-precision trained designs are fabricated.

Research Article Vol. 34, No. 1 / 12 Jan 2026 / Optics Express 1

Optics EXPRESS

QuATON: quantization aware training of optical neurons

HASINDU KARIYAWASAM,[1] RAMITH HETTIARACHCHI,[1] QUANSAN YANG,[2,3,4] ALEX MATLOCK,[3] TAKAHIRO NAMBARA,[2,5] HIROYUKI KUSAKA,[2,5] YUICHIRO KUNAI,[2,5] PETER T. C. SO,[3,6] EDWARD S. BOYDEN,[2,6,7,8,9,10,11] AND DUSHAN N. WADDUWAGE[1,12,13,14,*]

[1]*Center for Advanced Imaging, Faculty of Arts and Sciences, Harvard University, Cambridge, MA 02138, USA*
[2]*McGovern Institute for Brain Research, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA*
[3]*Department of Mechanical Engineering, MIT, Cambridge, MA 02139, USA*
[4]*Department of Materials Science and Engineering, University of Washington, Seattle, WA 98195, USA*
[5]*Advanced Research Core, Fujikura Ltd., Kiba, Tokyo, Japan*
[6]*Department of Biological Engineering, MIT, Cambridge, MA 02139, USA*
[7]*Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA*
[8]*Howard Hughes Medical Institute, MIT, Cambridge, MA 02139, USA*
[9]*K. Lisa Yang Center for Bionics, MIT, Cambridge, MA 02139, USA*
[10]*Center for Neurobiological Engineering, MIT, Cambridge, MA 02139, USA*
[11]*Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA 02139, USA*
[12]*School of Data Science, Old dominion university, Norfolk, VA 23529, USA*
[13]*Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA*
[14]*Department of Physics, Old Dominion University, Norfolk, VA 23529, USA*
[*]*dwadduwa@odu.edu*

Abstract: Optical processors, built with "optical neurons," can perform large-scale high-dimensional linear operations at the speed of light. With the current advances in micro-fabrication, such optical processors can now be 3D-fabricated, but at limited precision, eventually leading to a model mismatch due to quantized optical weights. To address this issue, we propose a quantization-aware training framework. Our approach accounts for physical constraints during the training process, leading to robust designs. We numerically demonstrate that our approach can design state-of-the-art optical processors using diffractive networks for multiple tasks despite quantized learnable parameters. We thus lay the foundation upon which improved optical processors may be 3D-fabricated in the future.

1. Introduction

Current advancements in deep learning have renewed interest in optical processors as a promising avenue to accelerate large-scale linear computations. While such optical accelerators have not yet been fully realized for arbitrary linear operations, optical neural processors (i.e., optical processors equivalent to neural networks) have already shown great promise in a variety of applications. For instance, diffractive deep neural networks (D2NN; plural: D2NNs) [1], made of a cascade of passive diffractive layers, could perform all-optical classification [1], all-optical quantitative phase imaging (QPI) [2,3], optical logic operations [4], spatiotemporal signal processing [5], saliency segmentation [6], and 3D object detection [7]. Similarly, learnable optical Fourier processors have shown to be capable of all-optical QPI [3], medical image processing [8,9], optical image encryption [10], and image classification [11]. Such optical processors have also been used as coding elements to design end-to-end optimized computational imaging systems [12,13]. All these optical processors linearly map an input light field to an output field. They thus

#546074 https://doi.org/10.1364/OE.546074
Journal © 2026 Received 1 Jul 2025; revised 3 Nov 2025; accepted 9 Dec 2025; published 2 Jan 2026

Optics EXPRESS

perform specific linear operations "learned" from training data, using computational elements we term "optical neurons."

Optical neurons are traditionally implemented using reconfigurable optical elements such as digital micromirror devices (DMDs) and spatial light modulators (SLMs). The state of each DMD micromirror or SLM pixel is treated as the learnable parameter of the optical neuron. However, recent advancements in micro-fabrication have now enabled the fabrication of 3D optics by voxel-wise modulating the refractive index of an optical substrate. Therefore, we now possess the unique capability to 3D print optical processors upon learnable design. Here, the transmission coefficient at each 3D location is treated as the learnable parameter of the optical neuron. In either case, however, —unlike the parameters of artificial neurons that can represent any real value— the parameters of optical neurons can only represent a limited set of complex values, constrained due to their physical characteristics and fabrication limitations. More precisely, parameters in optical neurons are quantized, complex-valued, and bounded. For example, typical SLM pixels enable 8-bit quantized phase-only parameters bounded in $[0, 2\pi)$; DMD micromirrors enable 1-bit quantized amplitude-only parameters bounded in $\{0, 1\}$. Similarly in 3D optics, fabrication limitations essentially constrain the parameters to a set of quantized complex values that are bounded in phase ($[0, 2\pi)$). These constraints may be ignored while designing (or training) the optical processor. But the change in the parameter distribution from design to fabrication, i.e., model mismatch, frequently leads to a performance decline in the realized system [14]. For example, recent work on all-optical QPI using D2NNs shows that reducing the precision of physical parameters exponentially decreases the performance of the D2NN [2]. Nevertheless, beyond such isolated examples, there has been little exploration [15] into how both *boundedness* and *quantization* of parameters affect the performance of optical neurons.

While quantization of optical neurons remains unexplored, quantization of parameters in artificial neural networks has been extensively studied in the machine learning literature to deploy models in resource constraint devices [16–18]. These works address, quantization post-hoc [19–21] or during the training –using quantization aware training (QAT) methods [18,22,23]. Post-hoc quantized models are easier to train but perform poorly during inference due to model mismatch; QAT models, on the other hand, are harder to train but perform faithfully to the trained model. The training difficulty lies in the quantization operation, which is not differentiable. This non-differentiability impedes gradient-based optimizers, making quantized models harder to train. QAT methods address this issue using few ways. The most-straightforward approach is to inject quantization noise during forward propagation such that the model learns to be robust; this technique has been used in optical neural networks to train neuromorphic models [24]. Another is to use the straight-through estimator (STE) [25,26] that passes the output gradient directly to the input during backpropagation. STE has been used to train optical neurons with limited precision controls and device variations [26]. A third approach is to use differentiable soft quantization functions [22,23] that approximate the quantized model. Soft quantization functions have not been effectively utilized in optical neurons, and are the focus of this work. Of note, however, Gumbel-Softmax (GS) –that enjoys widespread use in neural networks as a differentiable approximation to the categorical distribution– has been used as a QAT technique in optical neural networks [15]. While conceptually similar, GS is not a direct approximation of the quantization operation and hence is not as effective as soft quantization functions (see Fig. 1).

To this end, inspired by recent work in computer vision [22,23], we introduce "Quantization Aware Training of Optical Neurons", or "QuATON", a QAT framework specifically targeting optical neurons. The key elements of this framework consist of: (1) a soft quantization function constructed using shifted sigmoid functions that gradually evolve towards the desired hard quantization levels; (2) an auto-tuning temperature parameter to control the quantization function during training. In comprehensive performance comparisons, we show the superiority of

Optics EXPRESS

QuATON over competing methods used to train optical neurons (see Fig. 1). Our work lays the foundation upon which improved optical processors may be designed and built in the future.

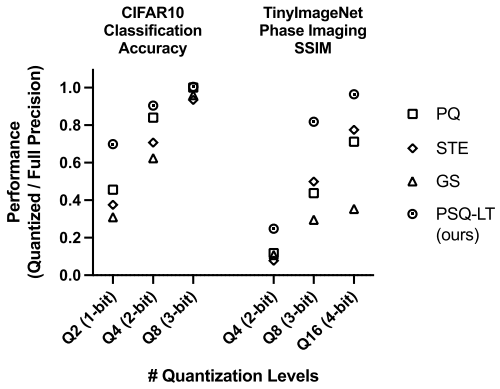

Fig. 1. Representative results for performance of QuATON (PSQ-LT) compared to other QAT methods (PQ, STE, GS) used to train optical neurons. Here PSQ-LT, PQ, STE, and GS respectively stands for progressive sigmoid quantization with learnable temperature, post quantization, straight-through estimator, and gumbel softmax.

2. Results and discussion

2.1. QuATON: quantization-aware training of optical neurons

The quantization operation maps a given continuous variable $x \in \mathbb{R}$, to a set of discrete values. These discrete values are known as quantization levels. In uniform quantization, they are evenly spaced and the uniform quantization operation is defined as

$$Q_h(x) = \begin{cases} l & \text{if } x<l \\ round\left(\frac{x-l}{\Delta}\right)\Delta + l & \text{if } l \le x<u \\ u & \text{if } x \ge u \end{cases}. \quad (1)$$

Here, $round(.)$ is the rounding operation to the nearest integer and $\Delta = \frac{u-l}{N-1}$ is the step size. $l, u \in \mathbb{R}$ are the lower and upper bounds of the quantized range, where $N \in \mathbb{Z}$ is the number of quantization levels. We hereon refer to the operation described in Eq. (1) as the *hard-quantization function*.

Due to its step-like nature, the hard-quantization function (Eq. (1)) has a zero derivative almost everywhere except for the sharp transitions between two quantization levels, where its derivative is undefined. Gradient-based optimizers –which depend on the derivative– thus fail in the presence of the hard-quantization function. In artificial- or deep- neural networks, the issues of hard-quantization have been addressed using soft-quantization functions. For instance, differentiable soft quantization (DSQ) [23] is a soft-quantization function based on the hyperbolic tangent (tanh). It has a learnable temperature parameter which controls the steepness of the transitions between two quantization levels. The temperature is updated during the training process, transforming the DSQ function closer to hard-quantization. Inspired by this, we developed a separate soft-quantization function named *Progressive Sigmoid Quantization (PSQ)* to train optical neurons aware of quantization. PSQ is different from DSQ in two ways. First, PSQ is based on the sigmoid function. Second, in DSQ, the weights outside the quantization range are clamped to the lower and upper bounds of the range. This results in a zero gradient for

Optics EXPRESS

the weights outside the quantization range, causing them to freeze during training. In PSQ, we removed this clamping, which provides a non-zero gradient to weights outside the quantization range.

The proposed progressive sigmoid quantization (PSQ) function is defined as,

$$Q_s(x,\tau) = l + \sum_{i=0}^{N-2} \Delta sig\left(\tau\left(x - l - \frac{\Delta}{2} - i\Delta\right)\right), \tag{2}$$

where

$$sig(x) = \frac{1}{1+e^{-x}}, \tag{3}$$

and τ is the temperature factor that changes the steepness of transition between two adjacent quantization levels (other variables are the same as in Eq. (1)). As τ increases, the PSQ function approaches the hard-quantization function, as shown in Fig. 2-D. When using PSQ in QuATON, we start training with a smaller τ value (i.e., a relaxed quantization function). We then gradually increase τ as the training progresses until PSQ (Q_s in Eq. (2)) approaches the hard-quantization function (Q_h in Eq. (1)). We call this *progressive training*. In this work, we consider two progressive training approaches: 1) linearly increasing temperature (PSQ-LI); and 2) treating the temperature learnable (PSQ-LT). In PSQ-LI, we start with a small τ value and we increase it linearly with the number of epochs. In PSQ-LT, we use gradient-based optimization to adjust τ as the training progresses. These two schemes are described in detail in the Methods section (under Progressive Training Schemes).

2.2. QuATON in diffractive networks

In this section, we use QuATON to train D2NNs, a type of optical processor first demonstrated by Lin et al. for all-optical image processing at terahertz wavelengths [1]. D2NNs consist of diffractive layers through which light passes, acting as 2D fully-connected networks. Each layer has discrete spatial locations (neurons) with complex transmission coefficients that modulate the amplitude and phase of the incoming light. While terahertz D2NNs have millimeter-scale neurons, visible-wavelength D2NNs –preferred for optical imaging– require nanometer-scale neurons, which are challenging to fabricate at full precision. Fabrication imposes constraints on the precision and bounds of the transmission coefficients. Thus, here we used QuATON to train visible-range D2NNs with quantized parameters, and thereby relaxing the required fabrication precision. Microfabrication technologies, such as two-photon lithography, allow manipulating the refractive index of any given 3D location on optical substrates at diffraction-limited resolution. In the context of D2NNs, this capability translates to phase-only optical neurons. Thus, our study considers phase-only D2NNs, and we perform numerical simulations of quantization-aware training approaches.

As shown in Fig. 2-A, let $E_{in}[x,y]$ and $E_{out}[x,y]$ be the input and output light fields to the D2NN. At the output light field, a detector is placed to capture its intensity $I_{out}[x,y] = |E_{out}[x,y]|^2$. During forward propagation, the raw phase coefficients are sent through the PSQ (or DSQ) function to obtain the soft-quantized phase coefficients (Fig. 2-C). The incoming light field to the layer is then modulated with the soft-quantized phase coefficients. For the n^{th} layer, this is given as

$$E^n_{out}[x,y] = E^n_{in}[x,y] \exp\left(jQ_s\left(\varphi_n[x,y], \tau_n\right)\right), \tag{4}$$

where E^n_{in} and E^n_{out} are the fields immediately before and after the layer n. φ_n and τ_n are the phase coefficients and quantization temperature of layer n, respectively. During backpropagation, the partial derivative of the layer output with respect to phase coefficients is computed. Similarly, for the learnable temperature case, the partial derivatives with respect to k_n (see Eq. (7)) are also computed for the optimization of the temperature parameters. A detailed description explaining

Approach: The paper proposes QuATON, a quantization-aware training framework using a progressive sigmoid soft-quantization (PSQ) function with progressive or learnable temperature to train optical neurons under realistic quantization constraints.

Optics EXPRESS

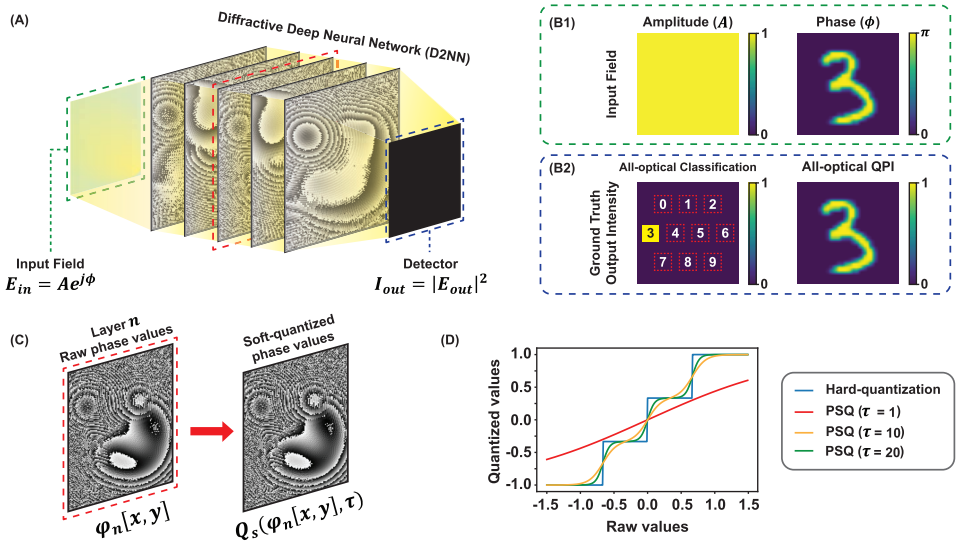

Fig. 2. Quantization-aware training of diffractive deep neural networks (D2NNs) using progressive sigmoid quantization (PSQ): A) D2NN architecture: The input field passes through the diffractive layers and the detector captures the output intensity. **B1)** an example of the amplitude and phase of the input field for the MNIST dataset. The input phase contains the information of interest. **B2)** the ground truth output intensities for the two tasks considered. For the all-optical classification task, the detector region is divided into 10 patches corresponding to each class shown in red dotted lines. For the example shown, the area corresponding to digit 3 is lighted up, and the other areas have zero intensity. For the all-optical quantitative phase imaging (QPI) task, the ground truth output intensity is proportional to the input phase. **C)** the training procedure optimizing only the phase coefficients of the D2NN. During forward propagation, the raw phase weights of the n^{th} layer ($\varphi_n[x, y]$) are sent through the PSQ function $Q_s(.)$. **D)** the evolution of the PSQ function with the temperature parameter (τ). When τ increases from 1 to 20, the function gradually becomes closer to hard quantization while keeping the differentiability.

the forward propagation and backpropagation through the D2NN is given in the Supplement 1 sec. A.

2.3. QuATON Designs State-of-the-art Diffractive Networks Despite Quantized Weights

As shown in Fig. 2-A and B, we designed D2NNs for two selected physics-based tasks: *all-optical classification* and *all-optical quantitative phase imaging (QPI)* [2,3]. In all-optical classification, the D2NN was trained to classify phase objects, while in all-optical QPI, it was trained to perform QPI of a phase object. For both tasks, the information of interest is in the phase of the input field. Therefore, the input light field to the D2NN models was constructed by placing images (from the datasets) in the phase after scaling to the range $[0, \pi]$. The amplitude of the input field was set to one throughout the field. An example of the input field to a D2NN is shown in Fig. 2-B1, and the field can be given as,

$$E_{in}[x, y] = e^{j\phi[x,y]}, \tag{5}$$

where $\phi[x, y]$ is the input phase, which is an image from the dataset scaled to $[0, \pi]$. The D2NN training details and the two physics-based tasks are explained in detail in the Methods section.

For each task, we experimented on multiple datasets. Our experiments were organized as follows: For each task and each dataset, we first trained D2NNs with full-precision weights

Introduced a quantization-aware training framework tailored for optical neurons.

Proposed Progressive Sigmoid Quantization (PSQ) with progressive and learnable temperature scheduling

Optics EXPRESS

without quantization. These models (denoted as FP) established a heuristic upper limit for the performance of a specific task on a particular dataset. We then hard-quantized the weights of the pre-trained full-precision models, generating results for post-quantization (PQ). We then trained our D2NNs using two existing QAT methods used to train optical neurons, straight-through estimator (STE) [25], and Gumbel-Softmax based quantization (GS) [15]. These experiments, i.e., PQ, STE, and GS, set the baseline of current state-of-the-art (SOTA). A detailed description of these methods is given in the Methods section. Finally, we trained D2NNs using QuATON. We experimented with four QuATON variations. We first used the same soft quantization and training mechanisms in differentiable soft quantization (DSQ) [23] (note that DSQ has not been used to train optical neurons before). We then used our proposed PSQ while keeping the temperature (τ) fixed (denoted as PSQ-FT). Last, we used PSQ with our two progressive training approaches, i.e. with a linearly increasing temperature (PSQ-LI) and the learnable temperature (PSQ-LT).

2.3.1. All-optical classification

We experimented on two classification datasets: MNIST [27] and CIFAR10 [28], as shown in Table 1. D2NN weights were quantized with 2, 4, and 8 quantization levels (Q2, Q4, and Q8). The full-precision D2NN achieved $\approx$ 90% accuracy for the MNIST dataset. For the more challenging CIFAR10 dataset, the classification accuracy was only 31.88%, even for the full-precision model. We observe that 8 quantization levels (Q8) were sufficient to achieve full-precision accuracy in both datasets. Even without quantization-aware training frameworks, with Q8, we achieved performance as good as the full-precision model. However, for 4 and 2 quantization levels (Q4 and Q2), existing methods (PQ, GS, and STE) showed degraded performance. Interestingly, at these levels, PQ performed better than GS and STE. This shows that the rounding error in PQ has yielded a better set of weights compared to the inferior gradients that steered the weights of STE and GS during training. In particular, the reason for PQ to yield better weights could be due to the classification task's simplicity, where the goal is to converge light towards a particular detector region. Compared to current methods, for Q4 and Q2, our QuATON variants (DSQ, PSQ-FT, PSQ-LI, and PSQ-LT) showed a clear improvement in performance. PSQ variants outperformed DSQ by a small margin across all datasets and quantization levels. The qualitative results are shown in Fig. 3. Of note, for Q2 in the MNIST dataset, QuATON methods showed more than a 50% increase in accuracy compared to current SOTA.

Table 1. Quantitative results of all-optical classification using D2NNs (classification accuracy)

Method	Proposed by	Used for Optical Neurons in	MNIST (Accuracy %)			CIFAR10 (Accuracy %)		
			Q2	Q4	Q8	Q2	Q4	Q8
FP	-	-		89.99			31.88	
PQ	-	[2]	21.84	86.89	90.06	14.54	26.79	31.89
STE	[25]	[26]	11.98	69.05	87.43	11.97	22.55	29.80
GS	[28]	[15]	12.79	55.19	83.50	9.86	19.88	30.62
DSQ	[23]	[a]	74.75	84.69	89.13	19.98	29.07	32.10
PSQ-FT	[a]	[a]	44.53	86.34	89.62	16.14	26.91	32.08
PSQ-LI	[a]	[a]	71.31	**87.73**	**90.08**	19.23	**30.18**	**32.35**
PSQ-LT	[a]	[a]	**75.03**	87.06	89.76	**22.27**	28.83	32.01

[a]This paper

Optics EXPRESS

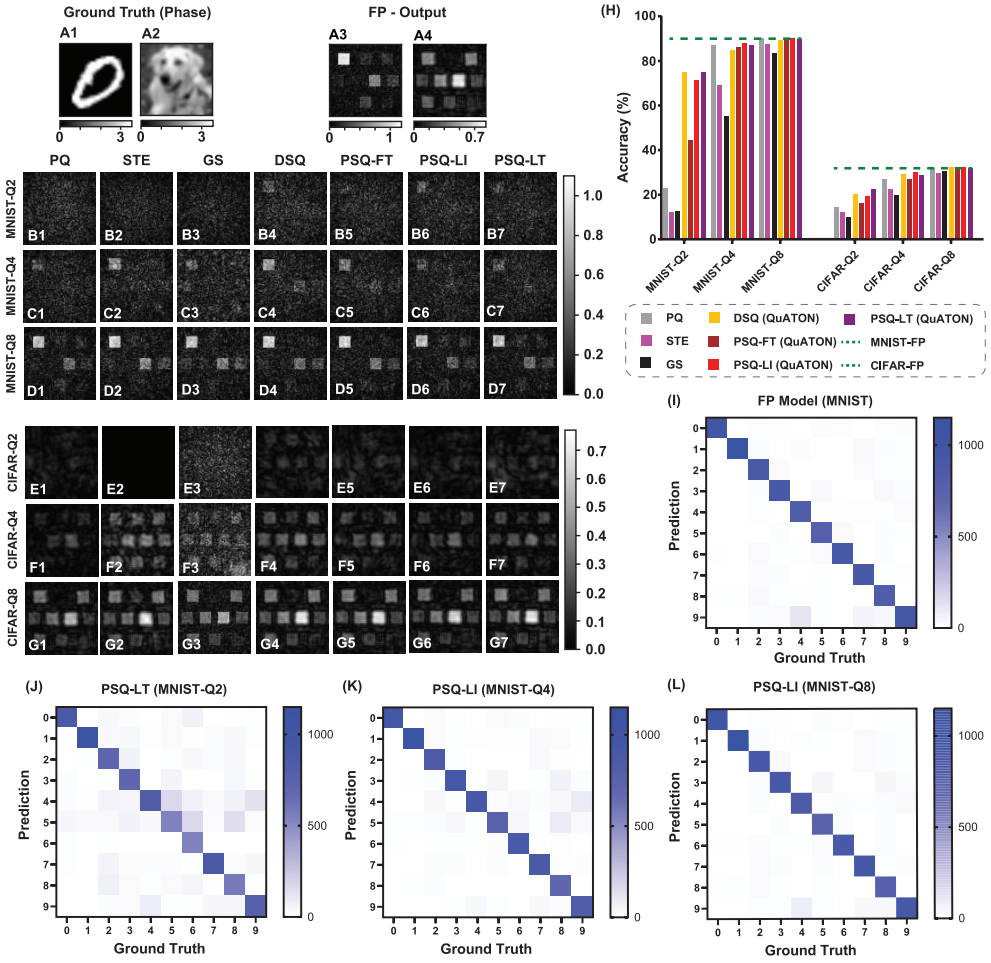

Fig. 3. All-optical classification results: A1–A2) Examples of input phase profiles for the two datasets. **A3–A4)** Output intensities of D2NNs trained with full-precision (FP) weights. **B–G)** Classification results of quantization-aware trained D2NNs for each example. Each row labeled ***x*-Q*n*** corresponds to dataset $\boldsymbol{x} \in \{\text{MNIST, CIFAR10}\}$, using D2NNs trained with $\boldsymbol{n}$-level quantized weights ($\boldsymbol{n} \in \{2, 4, 8\}$). Columns represent different QAT methods (indicated above row B). **H)** Quantitative comparison of classification accuracies. **I–L)** Confusion matrices for the FP model and the best-performing methods at each quantization level for MNIST.

2.3.2. All-optical quantitative phase imaging (QPI)

All-optical quantitative phase imaging translates the phase information of the input light field to the intensity at the output light field. It is thus an image translation task that's more challenging than the previous classification task. As shown in Table 2, we experimented on three datasets: MNIST, TinyImageNet [29], and RBC (red blood cells). D2NN weights were quantized with 4, 8, and 16 quantization levels (Q4, Q8, and Q16). The comparison of the quantitative results (using the structural similarity index measure - SSIM [30]) for this task is given in Table 2. The full-precision (FP) D2NN achieved an SSIM of 0.8560, 0.7385, and 0.9227 on the MNIST, TinyImageNet, and RBC datasets, respectively. Post-quantization of FP models severely degraded the model performance especially for Q4 and Q8 (see row PQ in Table 2). STE and GS

Optics EXPRESS

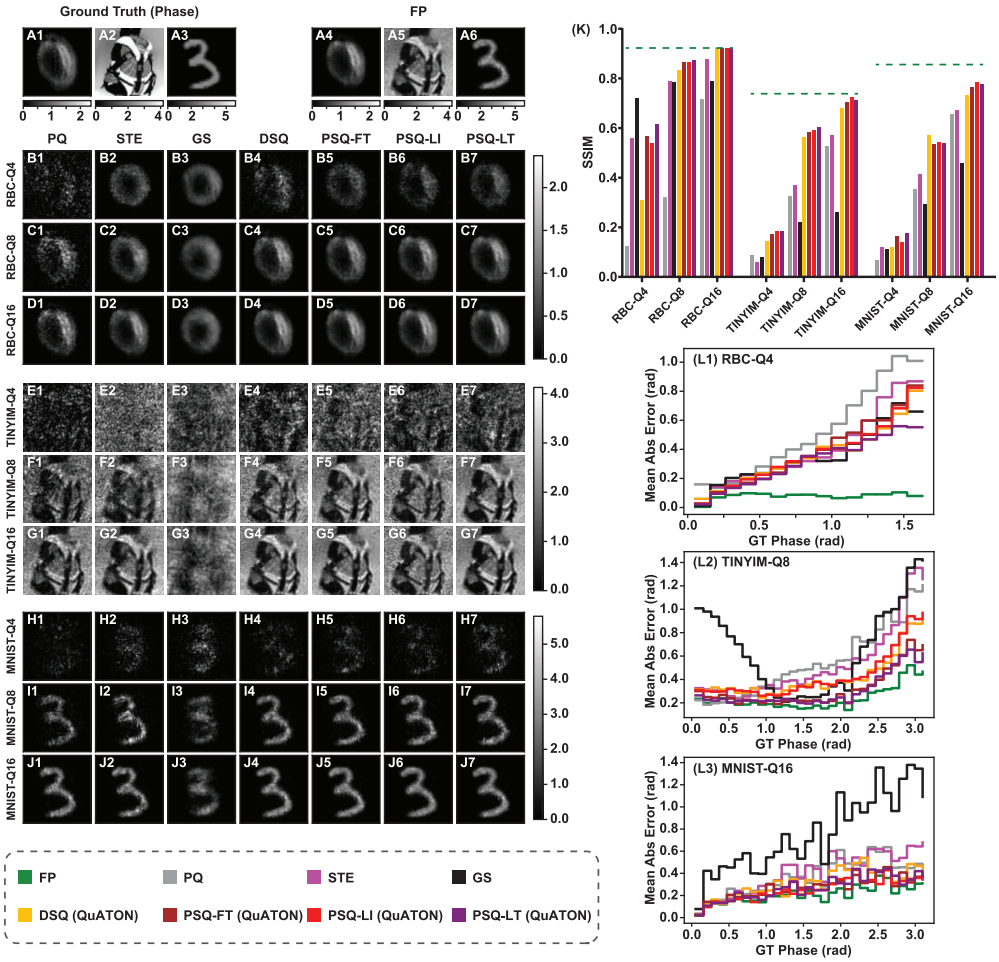

Fig. 4. All-optical quantitative phase imaging results: A1)-A3) Examples of the input phase for the three datasets. **A4-A6)** Corresponding output intensities ($\times\pi$) for full-precision (FP) D2NNs. **B)-J)** QPI results (output intensity $\times\,\pi$) for quantization-aware trained D2NNs. Each row ***x*-Q*n*** shows results for dataset $\boldsymbol{x} \in$ {RBC, TINYIM, MNIST} trained with $\boldsymbol{n}$-level quantized weights ($\boldsymbol{n} \in \{4, 8, 16\}$). Columns correspond to different QAT methods indicated above row B. **K)** Quantitative comparison (mean SSIM over test sets). **L1–L3)** Mean absolute phase error versus ground truth for **RBC-Q4**, **TINYIM-Q8**, and **MNIST-Q16**, shown for the examples above.

behaved differently on each dataset. On the MNIST and RBC datasets, STE improved the SSIM, while for the TinyImageNet dataset the improvement was not consistent. For the MNIST and TinyImageNet datasets, GS in fact degraded the SSIM values (compared to the PQ baseline), but surprisingly improved by a large margin for the RBC dataset. Our QuATON variants (DSQ, PSQ-FT, PSQ-LI, and PSQ-LT) clearly and consistently improved the SSIM across all datasets and quantization levels (compared to the PQ baseline). In almost all cases, at least one QuATON variant outperformed the current SOTA (i.e., PQ, STE and GS). Surprisingly, GS showed the highest SSIM for the RBC-Q4 case. For a deeper analysis of these results, we consider the qualitative results shown in Fig. 4.

Optics EXPRESS

Table 2. Quantitative results of all-optical quantitative phase imaging using D2NNs. The SSIM values are given for FP models and quantized models trained using different QAT methods

Method	MNIST			TinyImageNet			RBC		
	Q4	Q8	Q16	Q4	Q8	Q16	Q4	Q8	Q16
FP		0.8560			0.7385			0.9227	
PQ	0.0674	0.3526	0.6555	0.0870	0.3234	0.5258	0.1232	0.3214	0.7140
STE	0.1200	0.4138	0.6698	0.0570	0.3689	0.5721	0.5578	0.7898	0.8774
GS	0.1114	0.2938	0.4595	0.0793	0.2182	0.2610	**0.7188**	0.7854	0.7891
DSQ[a]	0.1207	**0.5701**	0.7321	0.1433	0.5610	0.6782	0.3095	0.8333	**0.9208**
PSQ-FT[a]	0.1653	0.5348	0.7627	0.1709	0.5845	0.7038	0.5678	0.8646	0.9199
PSQ-LI[a]	0.1411	0.5412	**0.7822**	**0.1854**	0.5924	**0.7237**	0.5394	0.8663	0.9206
PSQ-LT[a]	**0.1772**	0.5374	0.7759	0.1832	**0.6042**	0.7125	0.6156	**0.8723**	0.9202

[a]QuATON

Considering the qualitative results for the RBC dataset with 4-level quantization (RBC-Q4), although GS (Fig. 4-B3) resulted a higher SSIM value, it had failed to capture the morphological features of the red blood cell as well as the PSQ methods (Figs. 4-B6 and B7). Further analysis revealed that for the RBC dataset, the D2NNs trained with GS result in mode collapse. This is a failure mode where the model learns to generate a generic, "average" RBC image for any given input. In the challenging 4-level quantization case (RBC-Q4), this behavior leads to a deceptively high SSIM score, as the generic output is structurally similar to most images in the dataset (Fig. 4-B3 and Fig. S1. In contrast, other methods better captured the specific shape and orientation of individual cells, even if their reconstructed phase values were distorted, demonstrating that the raw SSIM metric does not convey the full story of reconstruction fidelity in this low-precision regime. This type of failure in the GS method can be attributed to its fundamental reliance on stochastic sampling. Unlike the direct, deterministic approximation used in our proposed method, this probabilistic approach can be unstable, causing the model to converge on a single, safe output mode rather than learning to reconstruct the full distribution of the data. Additional examples showing this are given in Supplement 1 Fig. S1. This is also evident from the mean phase error plot shown in Fig. 4-L1. It can be clearly seen that PSQ-LT has a lower mean phase error throughout the entire phase range, and especially for higher phase values compared to other methods for the considered example. For RBC-Q8 and RBC-Q16 (Fig. 4-C and D), our methods perform similarly to the FP model.

For TinyImageNet with 4-level quantization (TINYIM-Q4), none of the methods produces good results. However, it can be noted that our QuATON variants (Figs. 4-E4-E7) have managed to capture some of the features of the input phase. For the TINYIM-Q8 and TINYIM-Q16 cases, QuATON variants showed better performance than the current SOTA approaches. The qualitative results for the MNIST dataset also show similar trends as the TinyImageNet. However, for MNIST-Q4, PSQ has managed to capture the shape of the digit much better compared to other methods (Fig. 4-H). Figure 4-L3 shows the phase error plot for MNIST-Q16. This shows that similar to TinyImageNet, although PSQ and DSQ perform on par with each other, for certain phase ranges (e.g., from 1.5 rad to 2.5 rad) PSQ methods have lower errors than DSQ. For further validity of these observations, additional examples for each dataset are shown in the Supplement 1 Fig. S2. Furthermore, the progressive quantization of D2NN phase coefficients for the TINYIM-Q8 case using the PSQ-LT method is shown in Supplement 1 Fig. S3. Further analysis of the phase distributions in trained D2NN layers shows that QuATON variants preserve phase patterns closely matching those of the FP design while quantizing the phase values. Although PQ also exhibits a similar pattern to FP, the subtle differences in QuATON's phase

Optics EXPRESS

patterns and distributions lead to notable performance gains. In contrast, current SOTA methods (STE and GS) introduce significant distortions in the phase patterns, clamping many weights to the extremes of the phase range and thereby degrading performance relative to the FP design. The analysis corresponding to this is given in the Supplement 1 Fig. S4.

2.4. QuATON improves performance in multiple quantization regimes over existing methods

Taken together, QuATON's strength lies in its robust performance under realistic fabrication constraints. For simpler tasks such as all-optical classification, QuATON demonstrates a significant improvement at lower fabrication precision (≤ 4 quantization levels). Despite ever improving SLM and lithography technologies, this regime of improvement is still useful for both programmable and fabricated optical neurons. High-precision SLMs are either slow or limited in the number of addressable pixels; newer technologies like Ferroelectric liquid crystals (FLCs) overcomes some of these limitations at the cost of limited quantization levels. Moreover, well established amplitude-based digital micro-mirror devices (DMDs) -that can be alternatives to phase-based optical neurons- operate at two (or a few) quantization levels. On the fabrication front, while micrometer-sized neurons can be 3D fabricated with many quantization levels, fabricating precious nanometer-scale neurons in 3D with many quantization levels remains challenging. Thus, QuATON at ≤ 4 quantization levels remain critical to improve high-performance optical neurons.

For more complex tasks, such as the all-optical QPI, QuATON outperforms the SOTA methods at all quantization levels. For example, as given in Table 2, for the TinyImageNet dataset even with 16 quantization levels, PQ (SSIM: 0.5258), STE (SSIM: 0.5721), and GS (SSIM:0.2610) show poor performance, while QuATON variants show a significant improvement in SSIM (maximum SSIM of 0.7237). Therefore, QuATON's ability to deliver high-performing designs despite quantized optical neurons is a key advantage to build more complex linear optical processors in the future.

3. Summary

In this study, we presented a quantization-aware training (QAT) framework for optical neurons called QuATON. QuATON is based on a soft differentiable quantization function and a progressive training approach. We demonstrated the results of this method on two physics-based tasks performed by diffractive deep neural networks (D2NNs).

Our results show that QuATON outperforms the post-quantization, straight-through estimator, and Gumbel-Softmax which are the current state-of-the-art(SOTA) QAT methods used for optical neurons. D2NNs with quantized weights trained using our method managed to achieve similar performance to a model with full-precision, using a smaller number of quantization levels (4 levels for all-optical classification and 16 levels for all-optical QPI). Furthermore, progressive training based methods outperformed fixed temperature based methods in almost all cases showing its importance in the framework. Altogether, this comprehensive evaluation of QAT methods for optical neural architectures sheds light on which method to choose to achieve a desired level of performance for a task while being constrained to a given precision of physical parameters. Although not demonstrated in this work, our method can easily be extended to non-uniform quantization by combining multiple PSQ functions. It is important to note that this study presents a numerical simulation of the performance of the D2NNs. The physical realization process of these networks can introduce other noise and artifacts that should be considered during the training process. In future work, we aim to include other types of noise in the training process and fabricate them to experimentally validate the results.

In conclusion, our QAT framework addresses the issue of lack of precision in fabrication methods, optical devices, and analog-to-digital/ digital-to-analog conversions. We believe

Optics EXPRESS

that this work lays the foundation upon which optical neurons can be physically realized for challenging vision applications in the future.

4. Materials and methods

4.1. Progressive training schemes

This section describes the two progressive training schemes used in QuATON: 1) linearly increasing temperature (PSQ-LI) and 2) learnable temperature (PSQ-LT).

4.1.1. Linearly increasing temperature

In this approach, we start by setting τ to a small value and increase it linearly over the training epochs. The scheduling of τ has three hyperparameters: *the initial value of* τ (τ_0), *the increment step size* ($\Delta\tau$), and *the interval between two consecutive increments* (Δt). Using these hyperparameters, the temperature at epoch t ($\tau(t)$) can be expressed as

$$\tau(t) = \tau_0 + \left\lfloor \frac{t}{\Delta t} \right\rfloor \Delta\tau. \tag{6}$$

4.1.2. Learnable temperature

In learnable temperature scheduling, we optimize τ through backpropagation. In this context, we define *quantization instances* (QIs), each having a separate PSQ function that is characterized by its own specific temperature factor. A QI can consist of either a single trainable parameter or a set of such parameters. Consider an example of an optical processor with M number of QIs. For the m^{th} quantization instance, the temperature (τ_m) is defined as

$$\tau_m = \frac{1}{|k_m| + \gamma}, \tag{7}$$

where k_m is a trainable parameter and γ is a constant that sets the upper bound of τ_m. To promote the progressive nature in the optimization process of τ_m, we introduce a regularization term in the loss computation. For the t^{th} training epoch, the regularization term R_t is given by

$$R_t(\mathbf{k}) = \lambda_1 s_t \left(\|\mathbf{k}\|_2^2 - \lambda_2^2 \right). \tag{8}$$

In this equation, $\mathbf{k} = [k_1, k_2, \ldots, k_M]^T$ is the vector containing the k_m parameters (Eq. (7)) of each QI in the model and λ_1, λ_2 are hyperparameters. s_t is a constant that is updated every β epochs as $s_t = 2^{\lfloor t/\beta \rfloor}$. This term forces the temperature to increase every β epochs while allowing it to be updated through backpropagation. The overall loss is computed as

$$\mathcal{L}(y, \hat{y}; \mathbf{\Theta}, \mathbf{k}) = \mathcal{L}_f(y, \hat{y}; \mathbf{\Theta}, \mathbf{k}) + R_t(\mathbf{k}) \tag{9}$$

where y is the ground truth, $\hat{y}$ is the predicted output, $\mathcal{L}_f$ is the loss function specific to the task, and $\mathbf{\Theta}$ is the set of trainable parameters of the optical processor. Note that the regularization term ($R_t(\mathbf{k})$) is added to the loss only when the temperature is learnable.

4.2. All-optical classification

We consider the all-optical classification of phase images using a D2NN. The task is evaluated on two datasets: MNIST digits and CIFAR10. Since both datasets have 10 classes, in this task, the output plane has 10 spatially separated patches, as shown in Fig. 2-B2. The patch with the highest mean intensity determines the class. During training, the objective is to concentrate

Optics EXPRESS

most of the light on the patch corresponding to the ground truth class and suppress the light that scatters to other patches. The loss function for this task is given in Eq. (10),

$$\mathcal{L}_f(Y, |E_{out}|^2; \varphi, \mathbf{k}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{SE}_i \times \left(1 - \frac{Y_i}{11}\right)\right) \tag{10}$$

where $\mathrm{SE}_i = \left(Y_i - |E_{out_i}|^2\right)^2$ is the squared error (SE) for the i-th pixel. Y is the label map corresponding to the ground truth digit. For example, as shown in Fig. 2, only the patch corresponding to digit 3 will equal 1 in the label map (Y) for digit 3. In essence, the loss function is a weighted mean squared error, where more weight is given to the non-target region to penalize light scattering to those areas.

4.3. All-optical quantitative phase imaging

In the all-optical QPI task, we consider three datasets: MNIST digits, TinyImageNet, and an experimentally collected red blood cell (RBC) dataset. In this task, the D2NN is trained with the objective

$$\min_{\boldsymbol{\varphi},\mathbf{k}} \mathcal{L}\left(|E_{out}[x,y]|^2, \frac{\phi[x,y]}{\pi}; \boldsymbol{\varphi}, \mathbf{k}\right) \tag{11}$$

such that the output intensity is proportional to the input phase. Here, $\boldsymbol{\varphi}$ denotes the phase coefficients of the D2NN, $\mathcal{L}$ is the overall loss, $E_{out}[x,y]$ is the output field of the D2NN, and $\mathbf{k}$ is the vector containing k parameters of each D2NN layer. The overall loss is computed according to Eq. (9) and reverse Huber loss [31] is used as the loss function $\mathcal{L}_f$.

4.4. Training details

In both tasks, only the phase coefficients of the D2NNs are optimized. During the training process, we apply PSQ for QAT of the phase coefficients. For a given task and a given dataset, we first train a D2NN with full-precision (FP) weights without quantization for 100 epochs (200 epochs for CIFAR10 dataset). Then, starting from FP weights, we apply PSQ to train the D2NN with quantized phase coefficients for another 100 epochs. Since we consider only the phase coefficients, they lie in the range $[0, 2\pi)$ rad. However, the trained phase coefficients of the FP model are unwrapped phase values, thus distributing them over several phase cycles. As a better initialization for the QAT process, we wrap the FP weights to bring them to the range $[0, 2\pi)$. This prevents weights outside of $[0, 2\pi)$ from being clamped to the lowest or highest quantization level at the beginning of the QAT process.

Furthermore, during QAT we limit the range of quantization levels to $[0, 1.99\pi]$ rad since 0 rad and 2π rad correspond to the same phase shift. For example, if we consider phase weights with 4-level quantization, the corresponding quantization levels will be $[0, 0.663\pi, 1.327\pi, 1.990\pi]$ rad. However, in the all-optical classification task, for 2-level quantization, we use the levels $[0, \pi]$. We evaluate three versions of PSQ and compare their performance for both tasks: 1) PSQ with keeping τ fixed (PSQ-FT), 2) PSQ with linearly increasing τ (PSQ-LI), and 3) PSQ with learnable τ (PSQ-LT).

For the all-optical classification task, we use 7-layer D2NNs trained with 2, 4, and 8 quantization levels for each case. For the all-optical QPI task, we train each D2NN with three different quantization levels (4, 8, and 16). For this task, we use 5-layer D2NNs for both TinyImageNet and RBC datasets, and 7-layer D2NNs for the MNIST dataset. Further details on D2NN specifications and the training process are included in the Supplement 1 sec. B.

Optics EXPRESS

4.5. Comparison of performance

We compare the performance of PSQ for the considered tasks with several QAT methods, the FP model, and the post-quantized FP model. For a fair comparison, we initialize the D2NN phase weights with FP weights before QAT using all methods.

4.5.1. Post-quantization (PQ)

In this, the phase weights of the FP model are directly quantized using Eq. (1).

4.5.2. Straight-through estimator (STE) [25]

In this method, Eq. (1) is used in the forward pass while during the backpropagation, gradients are computed as,

$$\frac{\partial Q_h}{\partial x} = 1 \quad \forall x. \tag{12}$$

where x is any trainable parameter in the network.

4.5.3. Gumbel-Softmax based quantization (GS)

This method uses Gumbel-Softmax as a soft quantization function, and this has been demonstrated as a QAT technique for D2NNs by Li et al. [15]. When training D2NNs using GS, we use the linear annealed temperature schedule (temperature starts from 50 and decreases by 0.5 each epoch), which they have mentioned in the paper for best-performing results.

4.5.4. Differentiable soft quantization (DSQ)

This method was proposed by Gong et al. [23] as a QAT method for DNNs. In their original work, DSQ is implemented with a learnable temperature and lower and upper bounds of the quantization range. However, since we consider the phase weights of D2NNs in the range $[0, 2\pi)$, we keep the lower and upper bounds of the quantization range fixed.

Funding. National Institute of Mental Health (R21-MH130067); National Institutes of Health (R01HL158102, R01GM160726); Howard Hughes Medical Institute; Center for Advanced Imaging at Harvard University; John Harvard Distinguished Science Fellowship Program; Fujikura Inc.; John Doerr; Lisa Yang.

Acknowledgements. This work was supported by the Center for Advanced Imaging at Harvard University (H. K., R. H., and D. N. W.), and NIH R21-MH130067-01 (D. N. W.). D. N. W. further acknowledges support from the John Harvard Distinguished Science Fellowship Program within the FAS Division of Science of Harvard University. Q. Y., P. T. S., and E. S. B. acknowledge support from Fujikura Inc. A. M. and P. T. S. also acknowledge support from NIH R01HL158102. E. S. B. further acknowledges the support from HHMI, John Doerr, and Lisa Yang.

Disclosures. The authors declare that they have no competing interests.

Author contributions statement. H. K. and R. H. developed the methods and conducted the numerical experiments; H. K., R. H., and D. N. W. analysed results; Q. Y., T. N., H. K. and Y. K. formulated the fabrication requirements for optical neurons; A. M. prepared the RBC dataset; P. T. S. advised Q. Y., T. N. and A. M.; Y. K. advised T. N. and H. K.; E .S. B. advised Q. Y.; D. N. W. advised H. K. and R. H. and supervised the project.

Data availability. Data underlying the results presented in this paper are available in Ref. [28] (MNIST), Ref. [29] (CIFAR10), and Ref. [30] (TinyImageNet). The red blood cell data are not publicly available at this time, but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

References

1. X. Lin, Y. Rivenson, N. T. Yardimci, *et al.*, "All-optical machine learning using diffractive deep neural networks," Science **361**(6406), 1004–1008 (2018).
2. D. Mengu and A. Ozcan, "All-optical phase recovery: diffractive computing for quantitative phase imaging," Adv. Opt. Mater. **10**(15), 2200281 (2022).
3. K. Herath, U. Haputhanthri, R. Hettiarachchi, *et al.*, "Differentiable microscopy designs an all optical quantitative phase microscope," arXiv (2022).

Optics EXPRESS

4. C. Qian, X. Lin, X. Lin, *et al.*, "Performing optical logic operations by a diffractive neural network," Light:Sci. Appl. **9**(1), 59 (2020).
5. J. Zhou, H. Pu, and J. Yan, "Spatiotemporal diffractive deep neural networks," Opt. Express **32**(2), 1864–1877 (2024).
6. T. Yan, J. Wu, T. Zhou, *et al.*, "Fourier-space diffractive deep neural network," Phys. Rev. Lett. **123**(2), 023901 (2019).
7. J. Shi, L. Zhou, T. Liu, *et al.*, "Multiple-view d 2 nns array: realizing robust 3d object recognition," Opt. Lett. **46**(14), 3388–3391 (2021).
8. C. S. Yelleswarapu, S.-R. Kothapalli, and D. Rao, "Optical fourier techniques for medical image processing and phase contrast imaging," Opt. Commun. **281**(7), 1876–1888 (2008).
9. A. Panchangam, K. Sastry, D. Rao, *et al.*, "Processing of medical images using real-time optical fourier processing," Med. Phys. **28**(1), 22–27 (2001).
10. S. Liu, Q. Mi, and B. Zhu, "Optical image encryption with multistage and multichannel fractional fourier-domain filtering," Opt. Lett. **26**(16), 1242–1244 (2001).
11. M. Miscuglio, Z. Hu, S. Li, *et al.*, "Massively parallel amplitude-only fourier neural network," Optica **7**(12), 1812–1819 (2020).
12. U. Haputhanthri, K. Herath, R. Hettiarachchi, *et al.*, "From Hours to Seconds: Towards 100x Faster Quantitative Phase Imaging via Differentiable Microscopy," arXiv (2022).
13. M. R. Kellman, E. Bostan, N. A. Repina, *et al.*, "Physics-Based Learned Design: Optimized Coded-Illumination for Quantitative Phase Imaging," IEEE Trans. Comput. Imaging **5**(3), 344–353 (2019).
14. C. A. Metzler, H. Ikoma, Y. Peng, *et al.*, "Deep Optics for Single-Shot High-Dynamic-Range Imaging," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2020).
15. Y. Li, R. Chen, W. Gao, *et al.*, "Physics-Aware Differentiable Discrete Codesign for Diffractive Optical Neural Networks," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design* ACM, New York, NY, USA, 2022), pp.1–9.
16. O. Zafrir, G. Boudoukh, P. Izsak, *et al.*, "Q8bert: Quantized 8bit bert," in *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, (IEEE, 2019), pp.36–39.
17. J. Wu, C. Leng, Y. Wang, *et al.*, "Quantized convolutional neural networks for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp.4820–4828.
18. S. K. Esser, J. L. McKinstry, D. Bablani, *et al.*, "Learned step size quantization," arXiv (2019).
19. Z. Liu, Y. Wang, K. Han, *et al.*, "Post-training quantization for vision transformer," Advances in Neural Information Processing Systems **34**, 28092–28103 (2021).
20. M. Nagel, R. A. Amjad, M. Van Baalen, *et al.*, "Up or down- adaptive rounding for post-training quantization," in *International Conference on Machine Learning*, (PMLR, 2020), pp.7197–7206.
21. Y. Nahshan, B. Chmiel, C. Baskin, *et al.*, "Loss aware post-training quantization," Mach Learn **110**(11-12), 3245–3262 (2021).
22. J. Yang, X. Shen, J. Xing, *et al.*, "Quantization Networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, 2019), pp.7300–7308.
23. R. Gong, X. Liu, S. Jiang, *et al.*, "Differentiable soft quantization: Bridging full-precision and low-bit neural networks," in *Proceedings of the IEEE International Conference on Computer Vision, vol. 2019-October* (2019).
24. M. Kirtas, A. Oikonomou, N. Passalis, *et al.*, "Quantization-aware training for low precision photonic neural networks," Neural Networks **155**, 561–573 (2022).
25. Y. Bengio, N. Léonard, A. Courville, *et al.*, "Estimating or propagating gradients through stochastic neurons for conditional computation," arXiv (2013).
26. J. Gu, Z. Zhao, C. Feng, *et al.*, "Roq: A noise-aware quantization scheme towards robust optical neural networks with low-bit controls," in *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*, (IEEE, 2020), pp.1586–1589.
27. Y. Lecun, L. Bottou, Y. Bengio, *et al.*, "Gradient-based learning applied to document recognition," Proc. IEEE **86**(11), 2278–2324 (1998).
28. A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. rep., University of Toronto (2009).
29. Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," CS 231N **7**, 3 (2015).
30. Z. Wang, A. C. Bovik, H. R. Sheikh, *et al.*, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. on Image Process. **13**(4), 600–612 (2004).
31. L. Zwald and S. Lambert-Lacroix, "The BerHu penalty and the grouped effect," arXiv (2012).