中国激光

片上集成光学神经网络综述(特邀)

符庭钊[1,4,5]，孙润[2,3]，黄禹尧[2,3]，张检发[1,4,5]，杨四刚[2,3]，朱志宏[1,4,5]，陈宏伟[2,3*]

[1]国防科技大学前沿交叉学科学院，湖南 长沙 410073；

[2]清华大学电子工程系，北京 100084；

[3]北京信息科学与技术国家研究中心，北京 100084；

[4]国防科技大学新型纳米光电信息材料与器件湖南省重点实验室，湖南 长沙 410073；

[5]国防科技大学南湖之光实验室，湖南 长沙 410073

摘要 光学神经网络是区别于冯·诺依曼计算架构的一种高性能新型计算范式，具有低延时、低功耗、大带宽以及并行信号处理等优势。片上集成是光学神经网络微型化发展的一种典型方式，近年来片上集成光学神经网络获得了学术界及工业界的广泛关注。对基于不同计算单元结构的片上集成光学神经网络的相关研究工作进行了梳理，并分析了其设计原理、实现方法及系统架构特征。同时结合国内外最新研究进展，进一步分析了片上集成光学神经网络在计算单元大规模拓展、可重构、非线性运算和实用化等方面面临的挑战及其未来发展趋势。

关键词 集成光学；光计算；光学神经网络；芯片；人工智能

中图分类号 O436 **文献标志码** A **DOI**：10.3788/CJL231227

1 引 言

神经网络是通过模仿人类大脑神经系统而建立的数学模型，人工神经元实现的功能与生物神经元中的树突、细胞体、轴突以及突触的功能类似。神经网络的发展依赖于强大的计算力资源、先进的智能算法以及丰富的数据资源等。随着人工智能时代的到来，以深度学习算法为代表的先进算法在大数据资源的驱动下发展迅猛，从而推动了神经网络在各个领域中的应用，如计算机视觉[1]、自然语言处理[2-3]、语音识别[4]、自动驾驶[5-6]以及生物医疗[7-9]等。在过去二十年中，半导体工艺技术的发展不仅降低了单个器件的功耗，同时极大程度提高了单位面积内集成的晶体管数量，使得基于冯·诺依曼架构和串行逻辑处理的中央处理器(CPU)的计算性能得到不断提高，同时也造就了一批性能优于CPU的硬件处理器，如基于并行处理的图形处理器(GPU)、现场可编程门阵列(FPGA)以及专用集成芯片(ASIC)等[10]。这些先进计算硬件在过去一段时间里极大地满足了神经网络在各个应用领域中的计算力需求。然而，随着未来社会智能化程度的不断提升，神经网络在处理复杂任务时需要消耗巨大的计算力资源，包括计算速度和计算能耗。现阶段半导体工艺技术的加工节点已经接近物理极限[10]，拥有极小尺寸的片上器件极易受量子隧穿和热效应的影响而难以正常工作，因此通过提高半导体工艺的加工精度来进一步获得高计算力资源的方法将难以持续。此时，寻找并探索新的计算范式具有重要意义。

神经网络工作过程中的大量矩阵运算可以用光的传播过程来等效。由于光的传播过程自身具有低延时、低功耗、大带宽以及并行信号处理等天然优势，故学术界和工业界的研究人员对光学神经网络(ONN)展开了深入研究。现阶段关于ONNs的研究工作主要包括空间衍射ONNs[11-23]、片上集成ONNs[24-70]以及基于其他体光学元件构建的非衍射型空间ONNs[71-73]等。ONNs继承了光传播过程中的所有优势，对该优势进行充分利用将有望突破未来人工智能时代发展过程中的算力瓶颈[74-76]。

本文主要针对片上集成ONNs的相关研究工作展开系统介绍，分析其设计原理、实现方法以及系统架构特征等。根据设计片上集成ONNs的基本单元结构对其进行分类，并分别对基于马赫-曾德尔干涉仪(MZI)、微环谐振腔(MRR)/波分系统(WDM)、亚波长衍射结构以及其他类型单元结构设计的片上集成ONNs依次展开论述。论述过程中对不同架构的片上集成ONNs的优劣进行对比、分析和总结。最后，对现阶段片上集成ONNs存在的问题展开讨论，并探讨其未来的发展趋势以及面临的挑战。

收稿日期：2023-09-21；**修回日期**：2023-10-18；**录用日期**：2023-10-24；**网络首发日期**：2023-10-31

基金项目：国家自然科学基金(62135009)

通信作者：*chenhw@tsinghua.edu.cn

2 片上集成光学神经网络近期进展

2.1 基于MZI干涉结构的片上光学神经网络

传统光学分束器和移相器可以通过级联等方式实现酉矩阵的运算功能。1994年，Reck等[77]利用3个分束器和3个移相器以级联的方式构建了任意3×3维度的酉矩阵。实际上，MZI也可以实现光学分束器和移相器的功能。2016年，Clements等[78]基于MZI级联的方式提出了矩阵分解的方法，对Reck等[77]提出的矩阵构建方式进行了优化，使得其设计只需要Reck等[77]设计方式一半的光学深度，并且光学损耗更低。同年，Ribeiro等[79]通过MZI级联的方式构建了4×4维度的矩阵，并且制造了集成光子芯片，该工作利用自适应算法实现了片上矩阵的重构功能。2017年，Shen等[25]在硅基芯片上集成了56个可编程MZIs，并以拓扑级联的方式在片上分步完成了4×4维度的光学矩阵乘积运算，如图1(a)所示，其中左边虚线框和右边虚线框区域的MZI结构分别被用来实现酉矩阵和对角阵的

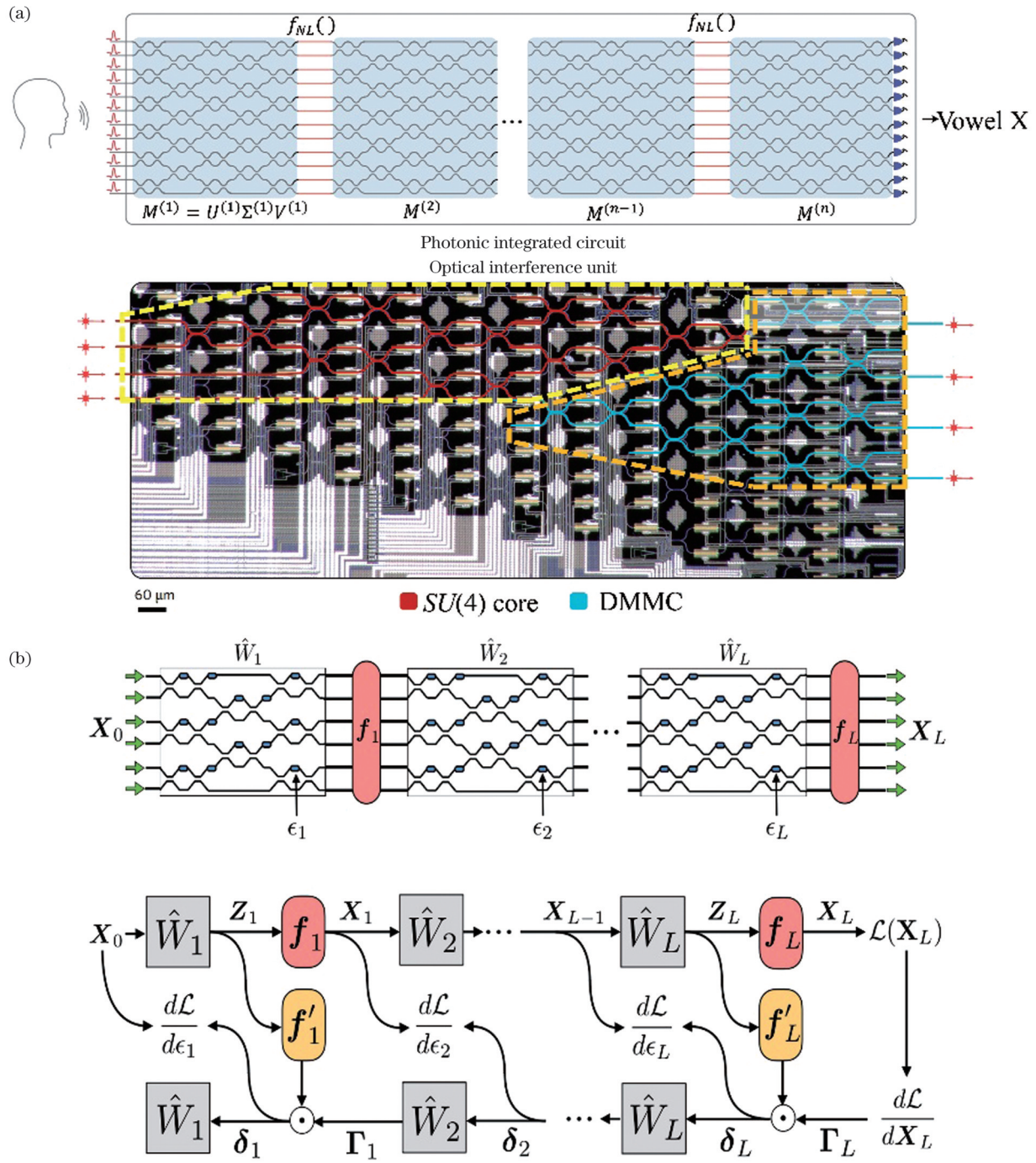

图1 基于MZI干涉结构的片上光学神经网络。(a)MZI拓扑级联阵列[25]；(b)支持原位在线训练及梯度反传的ONN[27]

Fig. 1 On-chip optical neural networks based on MZI interference structure. (a) MZI topology cascaded array[25]; (b) ONN supporting *in-situ* online training and gradient backpropagation[27]

运算功能。具体操作是通过奇异值分解的方法将任意矩阵($\boldsymbol{M}$)分解为一个对角阵($\boldsymbol{\Sigma}$)和两个酉矩阵($\boldsymbol{U}$和$\boldsymbol{V}$),即$\boldsymbol{M}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\dagger}$,其中符号“†”表示对矩阵$\boldsymbol{V}$求共轭。从理论分析可知:任何酉矩阵$\boldsymbol{U}$、$\boldsymbol{V}^{\dagger}$都可以通过光学分束器和移相器来实现,而对角阵$\boldsymbol{\Sigma}$可以通过光衰减器实现,这些功能都可以通过片上MZI拓扑级联的方式在物理器件上实现。此外,该工作基于元音识别数据集在实验上验证了片上集成ONN芯片的性能,该ONN芯片测试过程中引入的非线性层是通过在计算机上仿真饱和吸收体的光传输特性曲线来实现的。最后,该研究工作在实验测试过程中获得元音识别测试集的盲测准确率为76.7%。另外,Shen等[25]还提出了片上集成ONN芯片理论计算速度的评价方法:

$$R=2m\cdot N^{2}\cdot 10^{11}\cdot \mathrm{OPS}, \tag{1}$$

式中:N表示每一层上的节点数量;m表示系统中$N\times N$维度的矩阵数量;10^{11}表示本次计算过程中所用探测器的带宽为100 GHz;OPS表示每秒完成的操作次数;系数2表示每个节点在运算过程中同时完成了一次乘法和加法操作。

2018年,如图1(b)所示,Hughes等[27]基于MZI级联的方式提出了一种原位训练方法来优化ONNs的结构参数,利用伴随变量法实现ONNs的梯度求导,并通过测量ONNs芯片输出端口的光功率来精确地获得ONNs训练过程中的反传梯度。该工作直接在集成光芯片硬件上完成了ONNs的结构参数训练,在一定程度上加快了ONNs结构参数优化迭代的速度,同时全光训练过程的功耗更低,这使得ONNs的训练变得更加高效。该方法还具备以下优势:当片上集成ONNs芯片的结构参数优化完成后即可工作,无需进行额外的工艺误差和系统误差校准。以上基于MZIs拓扑级联方式展开的片上集成ONNs的相关研究工作都是基于光的强度(幅度)调制进行的。然而,对于光而言,同时考虑光的多个特征量进行ONNs结构参数训练将会使该网络在训练过程中拥有更高的自由度,且训练得到的ONNs也将会具有更强的逻辑推理能力。

2021年,Zhang等[41]基于MZIs级联的方式设计了一种复数值ONNs,如图2(a)所示,该工作在ONNs结构参数的训练过程中同时利用了光的相位和幅值,使得ONNs网络训练过程中的可调变量增加了一倍。该设计方法得到的ONNs性能更好,极大地提高了ONNs的计算速度和能效。另外,该复数值ONNs拥有更强的学习能力,包括分类精度以及损失函数的收敛速度等都比实值神经网络表现更好。该工作利用复数值ONNs分别完成了逻辑门运算、IRIS数据集类别预测、非线性数据(圆和螺旋)分类以及MNIST数据集手写数字识别等任务,并且均取得了较好的结果。其中,在MNIST测试集分类任务中,在相同矩阵规模下复数值ONN实现了90.5%的测试准确率,比实值神经网络的测试准确率高8.5%。

随着应用场景复杂程度的提升,大规模、高集成度以及低功耗的硬件对于ONNs而言变得越来越重要。2022年,Zhu等[47]在MZIs级联中引入了衍射单元,如图2(b)所示,利用这些衍射单元可以完成傅里叶变换及其逆变换,从而提升了ONNs在运算过程中处理的矩阵维度,同时降低了其计算能耗。传统ONNs设计框架中面积和能耗往往与输入矩阵的维度呈二次幂关系,而基于MZIs和衍射单元联合设计后,ONNs芯片的面积和能耗将与输入矩阵的维度呈线性关系,因此集成度和能耗均有所改善。这种新的设计方案与该课题组之前的工作[41]相比较,在相同的MNIST和Fashion-MNIST数据集的分类实验中,整个ONN的面积以及能耗约为之前仅基于MZIs拓扑级联方式实现的ONNs的1/10。

2.2 基于MRR波分结构的片上光学神经网络

MRR具有典型的滤波功能,可以筛选不同频率的光,并且可以对所选频率下的光功率进行再分配。Tait等[80-81]对MRR权重分配相关的研究工作进行了全面系统的分析和总结。基于此,MRR波分系统能够实现权值分配及求和运算功能,通过对滤波后的不同波长的权值进行巧妙设计就能够实现ONNs推理过程中的矩阵运算功能。

2019年,Feldmann等[31]基于MRR波分系统及相变材料(PCM)设计了一种全光学的神经突触系统,如图3(a)所示,该系统处理信息的方式与人的大脑更加类似。该方法实现的ONNs在工作过程中使用PCM单元(面积大小约为3.6 μm^{2})对输入脉冲进行加权处理,然后通过微环将相应波长的光耦合进单模波导中进行功率叠加,当单模波导中的光功率累积超过某个阈值时,最后一个MRR上的PCM单元将切换晶体状态并产生输出脉冲,至此在光上完成一次计算。该计算过程中最后一个MRR上的PCM单元在物理过程中实现了光学非线性功能,该光学非线性功能是通过调控PCM单元的晶体状态实现的。具体是当输入光功率的积分和低于一定阈值时,该PCM单元处于晶体状态,此时大量光功率被吸收,传播光无法通过;反之,当输入光功率积分和高于一定阈值时,PCM单元处于非晶体状态,此时大部分光功率被释放,即传播光可以通过。该研究工作设计并制造了由4个神经元和60个突触共140个片上光学元件组成的ONN,并基于模式识别等任务验证了该片上集成ONN的性能。2021年,Feldmann等[42]基于MRR波分系统和PCM材料在氮化硅(Si_3N_4)平台上进一步设计并制备了集成光子张量核,在芯片上以光的方式实现了传统卷积核的并行处理功能,它能够以每秒数万亿次操作的速度运行(即每秒完成10^{12}次乘法和加法

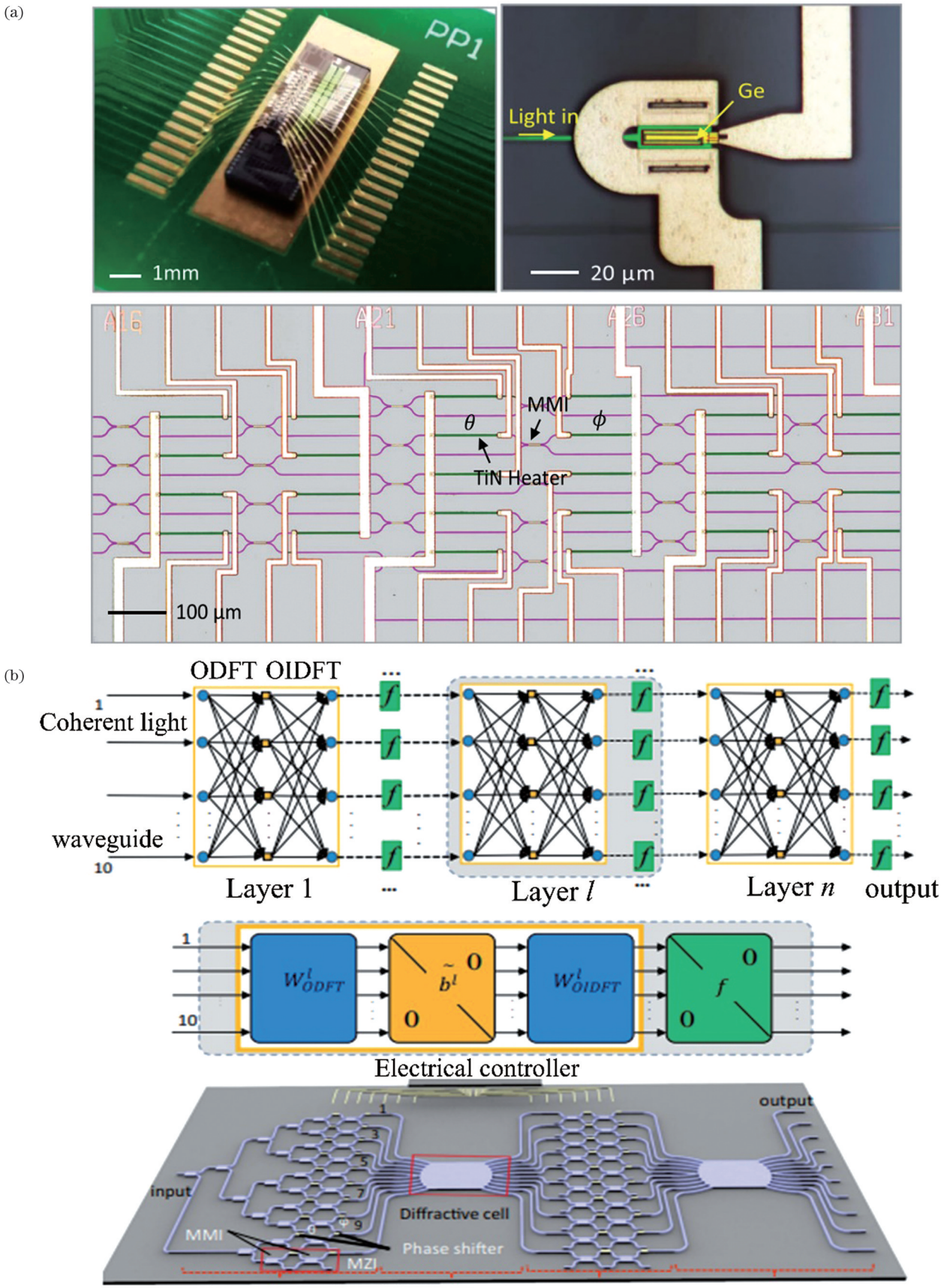

图 2 基于 MZI 干涉结构改进后的片上光学神经网络。(a)基于 MZI 阵列的复数值 ONN[41];(b)基于 MZI 阵列及衍射单元的 ONN[47]

Fig. 2 Improved on-chip optical neural networks based on MZI interference structure. (a) Complex ONN based on MZI array[41]; (b) ONN based on MZI array and diffractive units[47]

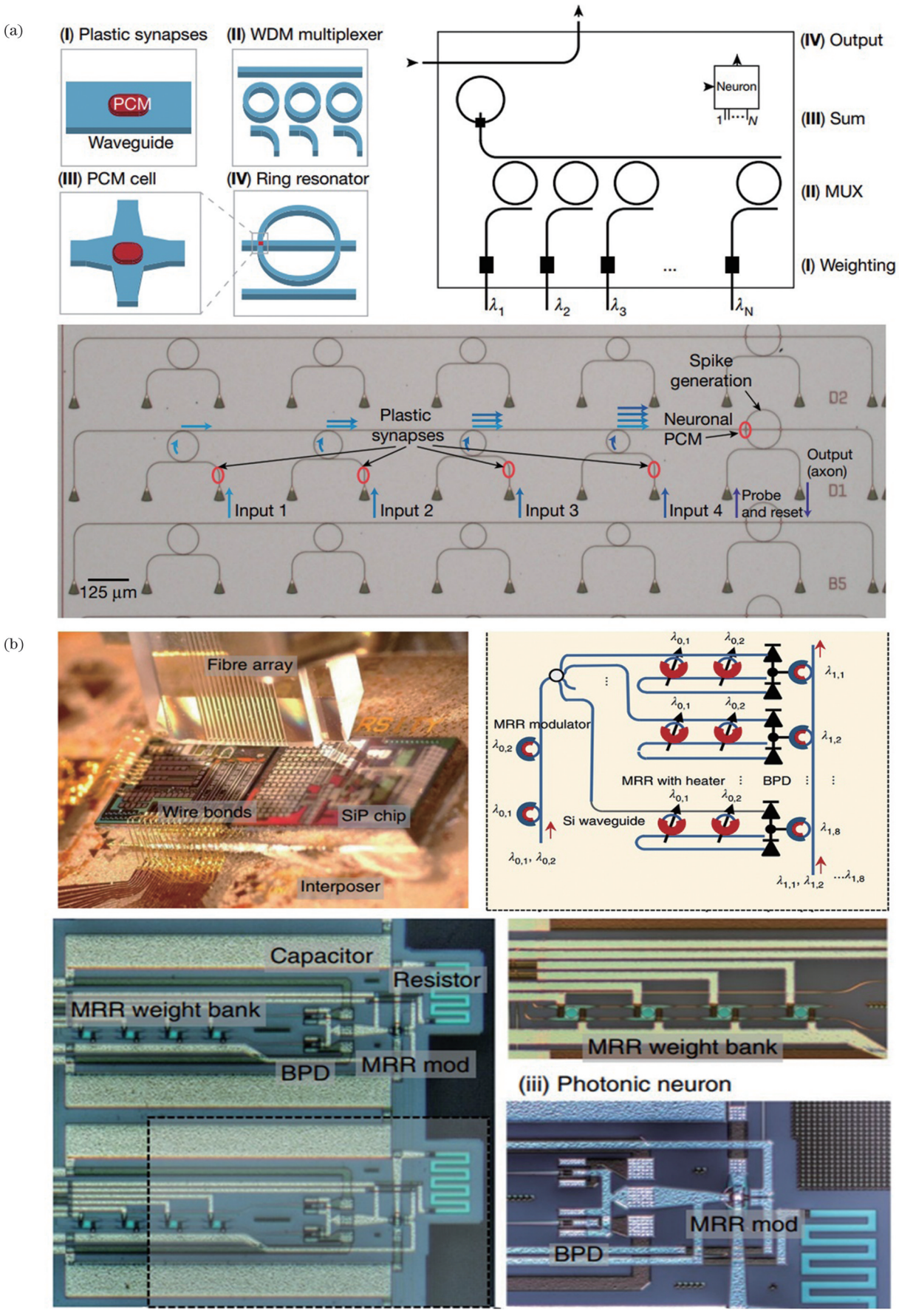

图 3 基于MRR波分结构的片上光学神经网络。(a)基于MRR波分系统及PCM单元的ONN[31];(b)用于光纤非线性补偿的ONN[43]

Fig. 3 On-chip optical neural networks based on MRR wavelength division structure. (a) ONN based on MRR wavelength division system and PCM units[31]; (b) ONN for fiber nonlinear compensation[43]

操作)。该工作利用PCM单元存储阵列,基于光子芯片的光频梳实现了并行光子内存计算,可以在超过14 GHz的带宽下工作,其运算速度仅受调制器和光电探测器速度的限制。另外,该团队设计并制备了一个片上光学卷积神经网络,当卷积核大小为2×2时,该ONN芯片对MNIST测试集的实验测试准确率达到了95.3%。

在光通信系统中,光纤的非线性和色散问题是提高远距离传输容量的主要障碍。针对光纤传输过程中的非线性问题,2021年,Huang等[43]基于MRR阵列在绝缘体上硅(SOI)平台上设计了一种用于解决海底光纤链路传输系统中光纤非线性补偿问题的片上集成ONN,如图3(b)所示。该可重构波分复用ONN系统能够在模拟域处理光信号,从而大大降低了传统数字信号处理电路对复杂度和速率的要求,且在实验上验证了该ONN系统能够在10080 km海底光纤链路通信系统中提高信号质量(*Q*)因子。2022年,Ohno等[57]基于MRR交叉阵列设计了一种可在线训练片上集成ONNs结构参数的硅基可编程光子集成回路(PICs),如图4(a)所示,该可编程PICs在执行矩阵向量乘法的过程中无需奇异值分解等复杂算法。该研究工作基于4×4维MRR交叉阵列设计并制备了一个具有三层网络的集成ONN芯片,并利用Iris数据集对该ONN芯片的性能进行了实验验证,获得了93%的识别准确率。值得强调的是基于MRR交叉阵列方案设计的片上集成ONNs支持梯度反向传播,文中以仿真的形式完成了片上反向传播的可行性验证,该项研究工作对于ONNs芯片结构参数的在线训练以及片上系统的误差校准具有一定的指导意义。2023年,Bai等[59]利用片上光频梳、MRR阵列以及螺旋硅波导延迟线设计了一个片上ONN系统。其中,光频梳由片上分布反馈式(DFB)激光器和一个MRR构成,如图4(b)所示。光频梳产生的多波长光,首先会经过一个MZI电光调制器,该电光调制器的作用是将输入信号加载到多波长上。完成信号加载后,多波长光信号将进入MRR阵列和螺旋硅波导延迟线阵列,此时每个特殊设计的MRR将会对不同波长进行筛选,同时对其完成幅度调制。最后,不同波长的光携带相应MRR加权后的光功率将在某一特定时刻被光电探测器接收到,如此反复工作即可完成片上集成ONNs运算过程中的所有矩阵向量乘法(卷积)操作。该项研究工作的特点在于将多波长源、数据加载区以及数据处理中心全部集成在同一芯片上,利用时间-波长拉伸的方法实现了卷积功能,其计算密度可达每平方毫米约1.04 TOPS(trillions of operations per second),即每平方毫米每秒完成约1万亿次操作。该计算密度与之前所有基于MRR波分系统设计的片上集成ONNs相比,其计算密度至少提高了5倍。另外,该研究工作中所设计的卷积神经网络在图像边缘质量检测及MNIST数据精度识别等方面与现阶段计算机的检测和识别结果相当。

2.3 基于亚波长衍射结构的片上光学神经网络

亚波长衍射结构属于超表面结构的一种,理论上而言,通过对超表面结构的巧妙设计几乎可以实现对反射或折射光束波前的任意调控[82]。在片上集成衍射光学神经网络(DONN)相关研究工作开展之前,Zhang等[83]和Wang等[84]分别在2017年和2019年基于SOI平台对片上超透镜相关工作进行了仿真和实验研究。

2020年,如图5(a)所示,Zarei等[35]基于高对比度透射阵列超表面设计了片上集成DONN,其中一维超表面由亚波长衍射结构排列而成,阵列中亚波长衍射结构的排列周期为500 nm,每个亚波长衍射结构的宽度为140 nm,工作波长为1550 nm。该工作将一维超表面定义为片上DONN的隐藏层,而隐藏层上的每一个亚波长衍射结构被定义为一个神经元。在亚波长衍射结构宽度及其排列周期固定的情况下,通过改变每一个亚波长衍射结构的长度就可以实现对通过该亚波长阵列的光束波前的调控。Zarei等[35]对于片上集成DONN的设计作了如下假设:与亚波长衍射结构相关的输出相位仅取决于其自身的几何参数,而不受其相邻亚波长衍射结构的影响。2021年,清华大学陈宏伟课题组提出了一种片上集成二维空域电磁传播模型[44]。该模型能够近似表征光在平板波导中的传播规律,因此可以将该模型的数学表达式作为片上集成DONN结构参数训练过程中的前向传播模型,从而可以在计算机上采用梯度下降等先进算法快速、高效地训练得到DONN的结构参数。同时该课题组还提出了一种权值映射模型[44],如图5(b)所示。该模型将三个具有相同尺寸的亚波长衍射结构作为一组,将其定义为一个新的基本衍射单元并作为片上集成DONN中的基本计算单元。该课题组提出的权值映射模型也考虑了传播光入射至亚波长衍射结构组时的入射角问题,因此可以保证预先训练获得的片上集成DONN结构参数能够更准确地在物理器件上实现。该研究工作基于加州大学欧文分校提供的心脏病患者数据集分别设计了三个片上集成DONN,这些DONNs分别包含了1、2、3个隐藏层,且其性能在仿真上均得到了有效验证。

2022年,Wang等[50]基于亚波长衍射结构设计并制备了片上集成DONN芯片,如图6(a)所示,并利用数字微镜器件(DMD)和透镜系统将输入信号加载到DONN芯片上,从而完成了片上集成DONN芯片性能的实验验证。为了减小相邻亚波长衍射结构单元间的相互干扰,Wang等[50]同样采用亚波长衍射结构组(将两个具有相同尺寸的亚波长衍射结构并为一组)作为

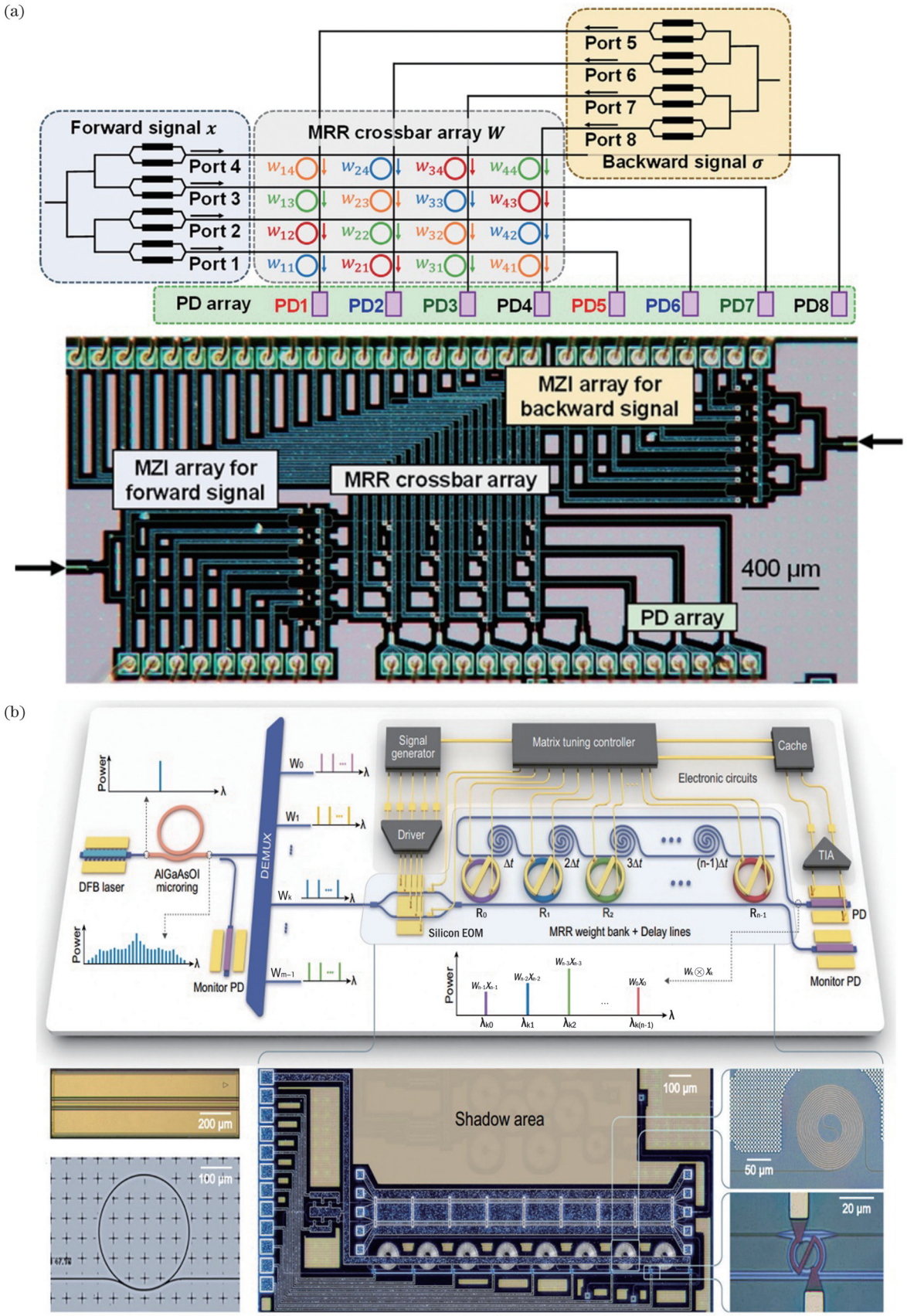

图 4　基于 MRRs 及其他辅助光学器件的片上光学神经网络。(a)基于 MRR 交叉阵列且支持梯度反向传播的 ONN[57]；(b)集光源、数据加载区以及数据处理单元于同一芯片上的 ONN 系统[59]

Fig. 4　On-chip optical neural networks based on MRRs and other auxiliary optical devices. (a) ONN based on MRR cross array and supporting gradient backpropagation[57]; (b) ONN system integrating light sources, data loading areas, and data processing units on single chip[59]

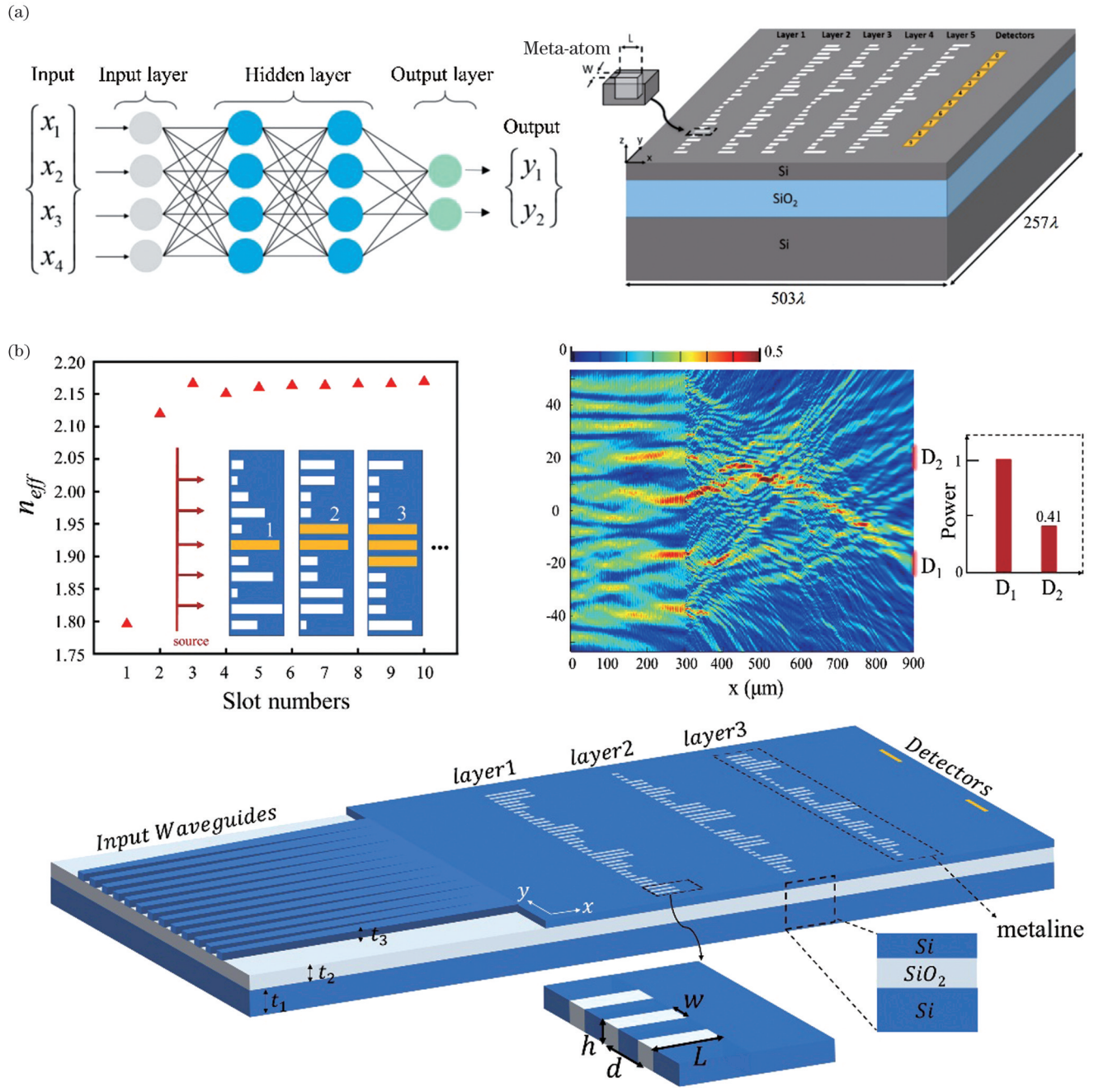

图 5 仿真验证的片上衍射光学神经网络。(a)计算单元由单个亚波长衍射结构表征的 DONN[35];(b)计算单元由亚波长衍射结构组表征的 DONN[44]

Fig. 5 On-chip diffractive optical neural networks verified by simulation. (a) DONN with computational unit characterized by single subwavelength diffractive structure [35]; (b) DONN with computational unit characterized by subwavelength diffractive structure group[44]

DONN 隐藏层中的一个计算单元。其中,亚波长衍射结构的排列周期为 500 nm,衍射结构组的排列周期为 1 μm,每个衍射结构的宽度均为 140 nm,这些结构参数沿用了该课题组在 2019 年研究片上超透镜时的相关参数[84]。2023 年,如图 6(b)所示,清华大学陈宏伟课题组在其前期的理论研究基础上,设计和制备了 DONN 芯片,并且在芯片制备过程中考虑了仿真时没有考虑的噪声干扰问题,如在片上 DONN 接收界面引入了片上光衰减结构以消除反射,避免片上集成 DONN 的推理结果受影响[60]。该研究工作为了有效地减小芯片制造和封装过程带来的系统误差,提出了一种可行的在线系统误差补偿方案,通过软硬件结合的方式,利用粒子群算法对封装完成后的 DONN 系统进行了在线误差补偿,该方案在很大程度上保证了 DONN 芯片的性能,同时也提高了实验测试系统的鲁棒性。

此外,2023 年,清华大学陈宏伟课题组提出采用拟合网络来实现片上集成 DONN 训练参数到芯片物

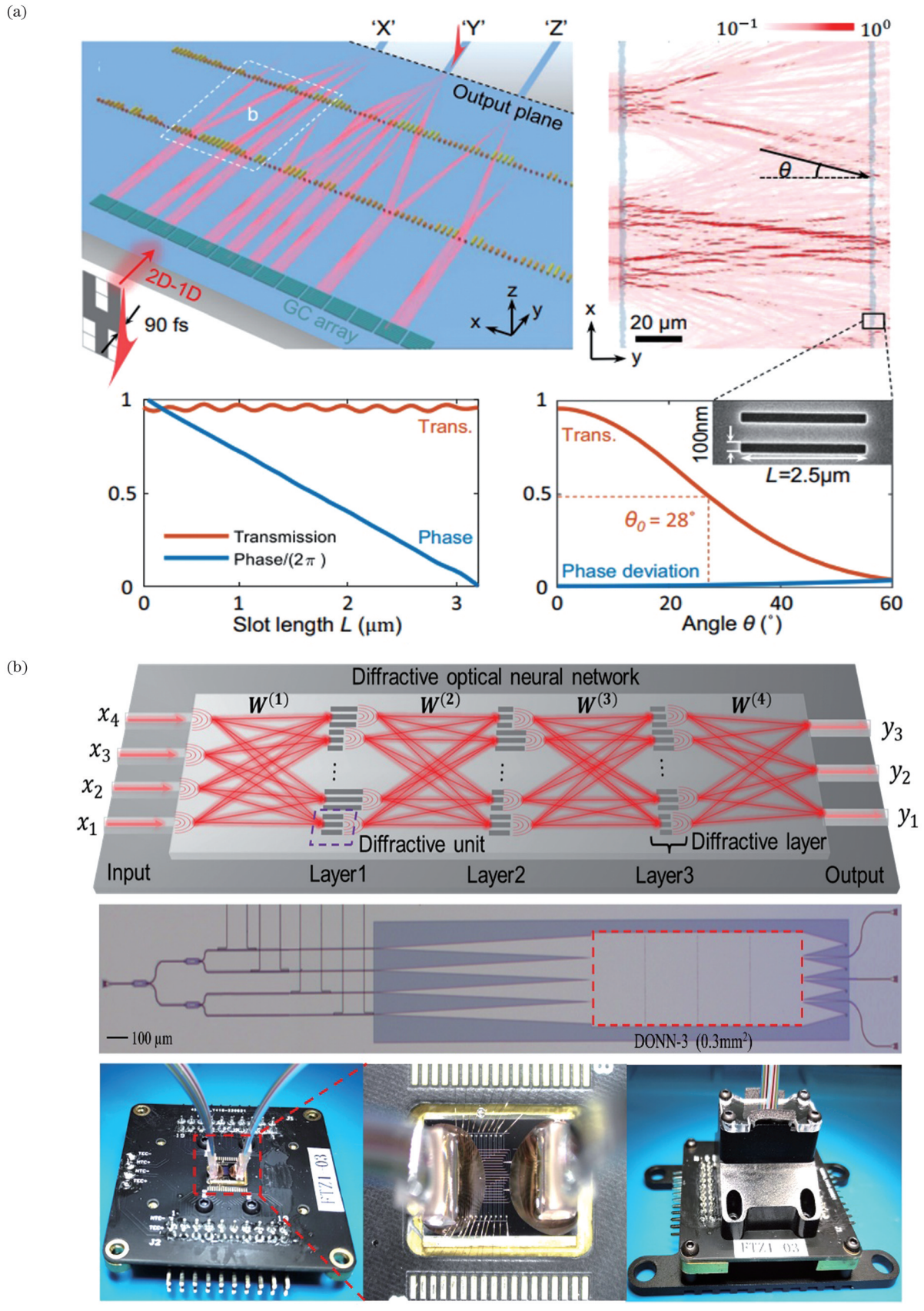

图 6 实验验证的片上衍射光学神经网络。(a)DONN 的计算单元由两个相同的亚波长衍射结构组成[50];(b)DONN 的计算单元由三个相同的亚波长衍射结构组成[60]

Fig. 6 On-chip diffractive optical neural networks verified by experiment. (a) DONN with computational unit composed by two identical subwavelength diffractive structures[50]; (b) DONN with computational unit composed by three identical subwavelength diffractive structures [60]

理结构的映射[65]。该方法相较于之前提出的权值映射模型，参数映射过程在很大程度上不再受权值映射模型近似条件的限制，从而大大提高了片上DONN计算单元的集成度，新映射规则下片上DONN在理论上可实现每平方毫米集成约6×10^4个计算单元。同年，该课题组在亚波长衍射结构单元的基础上设计了片上衍射光学卷积核，然后基于该衍射光学卷积核设计了片上光学卷积神经网络(OCNN)，并仿真验证了OCNNs在图像分类、图像去噪等方面的性能[64]。

2.4 其他结构的片上光学神经网络

本文中基于其他结构的片上ONNs是指基于片上二维集成波导、片上多模干涉仪(MMI)、片上移相器或衰减器以及片上三维集成波导等光学元器件的神经网络。

2020年，Moughames等[85]利用双光子聚合物打印技术制造了片上三维集成低损耗光子波导阵列，如图7(a)所示，其中光子波导的直径为1.2 μm，相邻波导的间距为20 μm，每立方毫米可集成约2200个神经元。同年，Qu等[36]利用逆向设计算法在SOI平台上设计并制备了硅基纳米光子散射单元，如图7(b)所示，通过逆向设计方法在4 μm×4 μm区域内制造了高自由度的纳米图案(纳米光子散射单元)。理论上，这些光学散射单元也可以作为片上集成ONNs的核心计算单元来实现任意酉矩阵的运算功能。基于

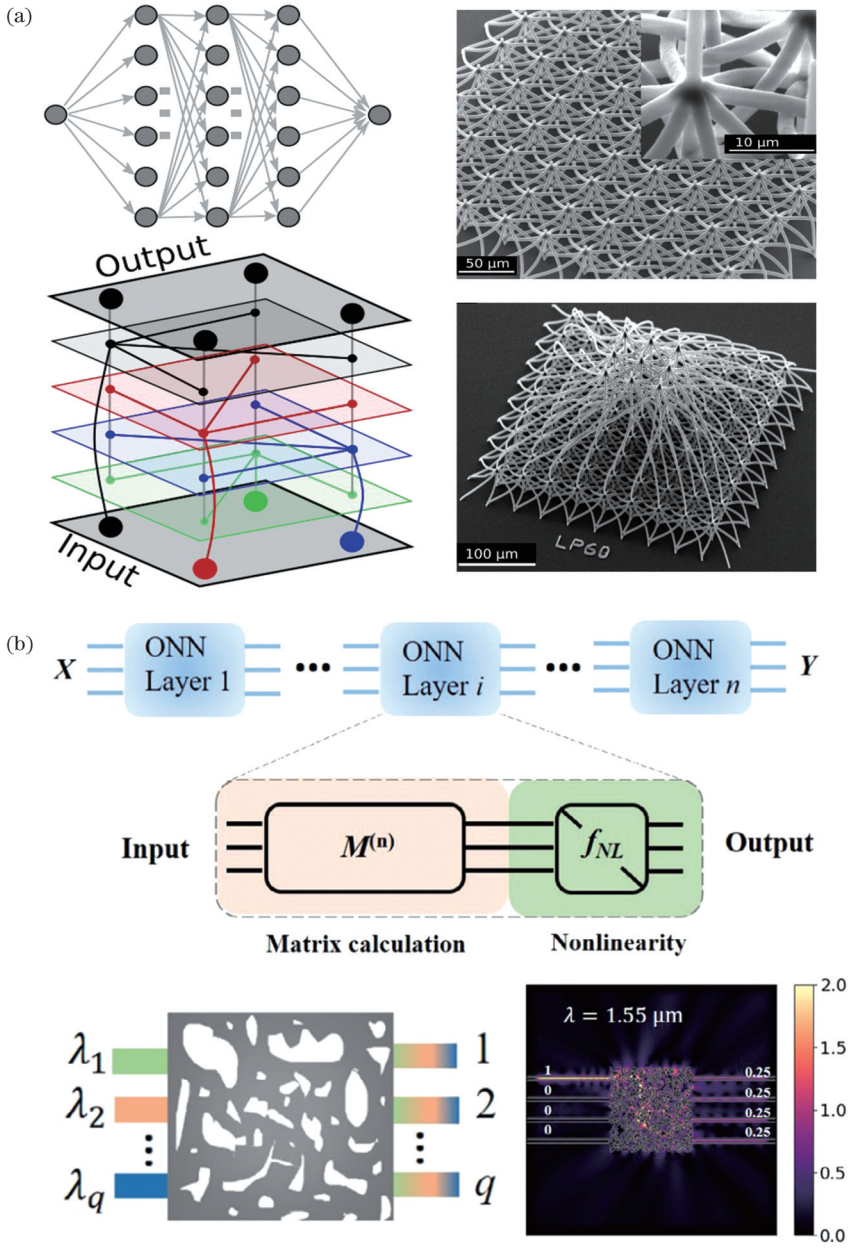

图7 基于其他结构的片上光学神经网络。(a)基于三维集成波导阵列的ONN[85]；(b)基于逆向设计方法设计的ONN[36]

Fig. 7 On-chip optical neural networks based on other structures. (a) ONN based on 3D integrated waveguide array[85]; (b) designed ONN based on inverse design method[36]

此，该团队针对 MNIST 数据集设计了片上集成 ONN，并在 MNIST 测试集上获得了 97.1% 的预测准确率。

2022 年，Ashtiani 等[53]利用载流子掺杂技术在硅波导上制造了电控可调衰减器，从而通过调节每根波导上衰减器的衰减系数来实现神经网络的权值配置，每根波导输出的光功率将会被多个光电探测器（PD）分别收集，然后这些光功率经过基于载流子注入设计的 MRR 调制器输出，该过程可以实现光学非线性功能，如图 8 所示。该硅基集成系统每次只能实现单个神经元的单次操作，如果要完成一个完整的矩阵运算则需要多次反复调制衰减器系数。例如，基于该架构设计一个包括输入、输出层在内的 3 层 ONNs，假设其输入层有 4 个输入节点，隐藏层有 3 个节点，输出层有 2 个输出节点，那么完成整个 ONNs 的矩阵运算就需要使该硬件系统在不同的衰减系数配置下至少循环 5 次以完成 4×3 和 3×2 两个矩阵的运算。

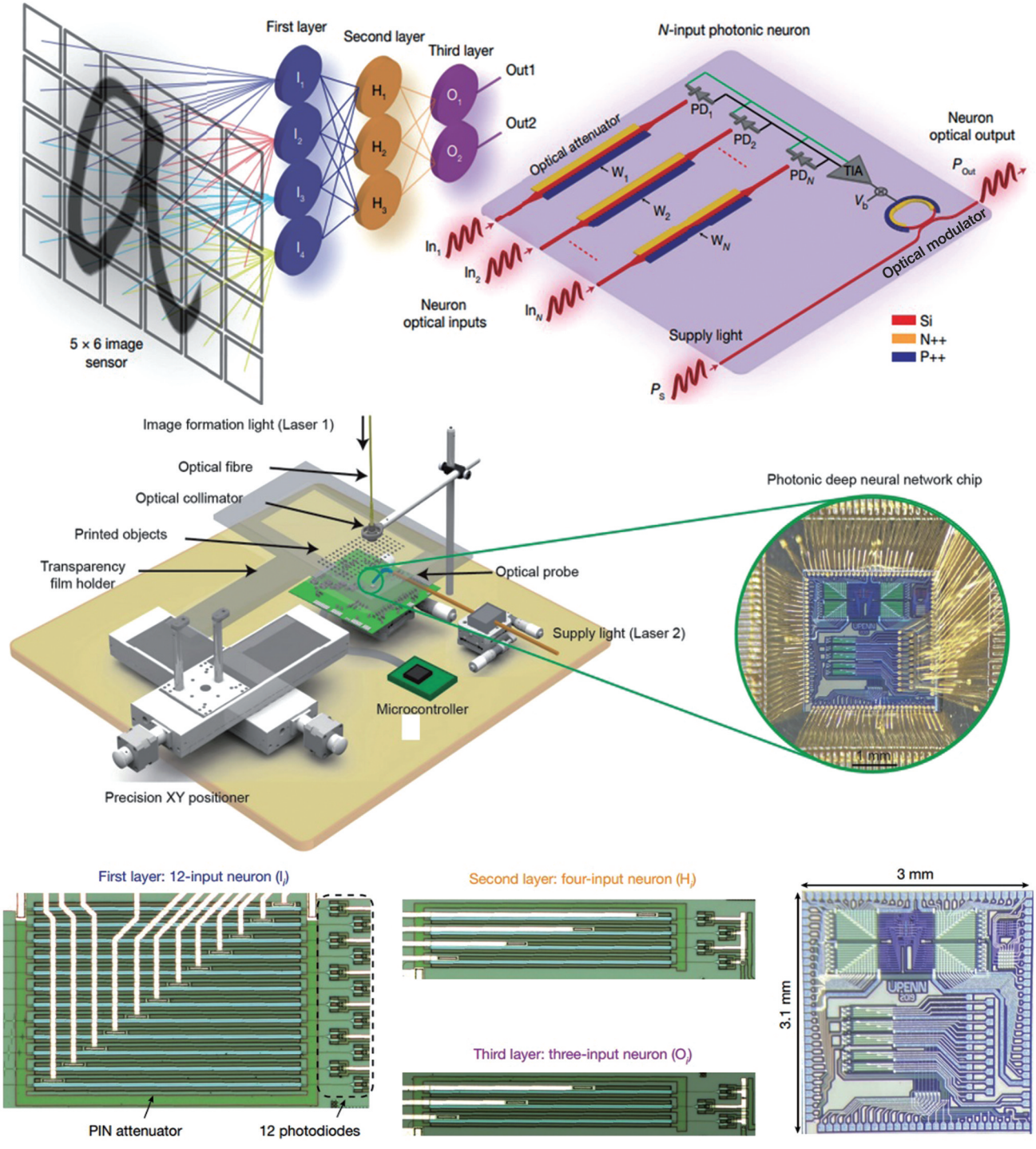

图 8　基于波导衰减调制器的 ONN[53]

Fig. 8　ONN based on waveguide attenuation modulators[53]

3 片上集成光学神经网络的研究现状分析与讨论

3.1 片上集成光学神经网络部分性能对比

综上所述，基于MZI、MRR等结构单元设计的片上集成ONNs均具备可重构功能，通过结合这些基本结构单元和PCM单元联合设计ONNs，就可以在片上实现非线性功能。然而，这类片上集成ONNs的基本结构单元（计算单元）的大规模拓展能力有限，计算密度低，能够并行处理的矩阵规模都较小。基于亚波长衍射结构单元设计的片上集成DONN，由于其计算单元尺寸小、集成度高、大规模拓展容易，故计算密度较高，并且具备并行处理大规模矩阵的能力。然而，正是因为片上集成DONN的计算单元尺寸较小，其在片上实现可重构、非线性功能时的难度比基于其他结构单元设计的ONNs更大。本文根据现阶段片上集成ONNs芯片的相关实验工作，在集成度、能耗以及计算吞吐量等方面对基于不同结构单元设计的ONNs进行了对比，具体情况如表1所示。其中：NBU/mm^{2}表示每平方毫米可集成的基本计算单元总数；J/operation表示完成每次操作所需要消耗的能量。

表1 片上集成ONNs的部分性能对比

Table 1 Partial performance comparison of on-chip integrated ONNs

Reference	Basic unit	Theoretical integration /(NBU /mm^2)	Operational power consumption /(J/operation)	Throughput /TOPS
Ref. [25]	MZI	＜10	7.66×10^{-14}	6.4
Ref. [41]	MZI	＜10	2.14×10^{-13}	21.6
Ref. [42]	MRR covered with PCM	＜5	5.9×10^{-15}	28.8
Ref. [47]	MZI and Diffractive cell	＜20	1.41×10^{-15}	32
Ref. [50]	Subwavelength unit	$\sim6.7\times10^{3}$	4.2×10^{-19}	4.05×10^{4}
Ref. [60]	Subwavelength unit	$\sim2\times10^{3}$	1.1×10^{-17}	1.38×10^{4}

表1中计算吞吐量采用的计算公式为本文中的式(1)。其中，关于NBU的对比，本文将参与片上集成ONNs矩阵运算的独立光学单元结构定义为一个基本计算单元。因此，基于MZIs设计的ONNs，其基本计算单元为MZI；基于MRRs设计的ONNs，其基本计算单元为MRR；基于亚波长衍射结构设计的DONN，在矩阵运算过程中，每个亚波长单元（SWU）以衍射方式参与计算。这些SWU的参数通过优化训练后以固定的周期排列构成DONN中的隐藏层，即每个隐藏层由不同的SWU排列组成，因此本文将SWU定义为片上集成DONN的基本计算单元。

3.2 片上集成光学神经网络未来发展的主要挑战

片上集成ONNs现阶段及未来主要面临的挑战包括以下几点：第一，在实现可重构调制功能的前提下，片上集成ONNs基本计算单元（如MZI等）的自身结构尺寸较大，计算单元在工作过程中需要额外的能量持续供给，且较多计算单元在高速调制下存在同步调制困难等问题。第二，引入非易失性PCM单元可以赋予ONNs可重构功能，且使得计算单元在工作过程中无需额外能量的持续供给。然而，新材料的引入在工艺制备过程中会带来新的难题，如插入损耗、PCM单元的重构速率以及如何建立统一的工艺标准等。第三，片上集成ONNs的深度难以进一步增加。首先，因为片上非线性的实现比较困难，故在ONNs中增加全连接层的数量意义不大。其次，即使解决了片上非线性难题，由非线性层引入的光功率损耗问题也将变得不可忽视，然而现阶段片上光功率放大机制也没有较好的实现方案，因此经过每个非线性层后光功率的衰减也会限制ONNs层数的规模。第四，片上并行信号的输入规模有限，虽然卷积神经网络可以实现输入信号的分块处理，但是分块处理意味着需要更多的光电转换处理，因此也在一定程度上降低了ONNs芯片的能效。第五，现阶段光电接口的速率限制了光芯片本身的优势，未来需要进一步提高光电接口的转换效率，包括速度和功耗。另外，光电转换过程中控制模块间的通信及控制协议也需要统一的标准，以使外围辅助电路网络的能耗更小。

3.3 片上集成光学神经网络未来发展的趋势

现阶段ONNs的发展分为两个方向，分别为全光神经网络和光电混合神经网络。全光神经网络发展目前遇到的阻碍较大，因为信号加载、信号存储、非线性激活等ONNs必备的基本功能在光上实现都比较困难。光电混合神经网络的发展阻力相对小些，因为光上难以实现的功能在电域上容易实现，然而电辅助的ONNs在计算速度及能效上与全光神经网络相比均有所降低，此时需要对光和电的计算力的分配进行权衡，最大化地发挥光和电混合计算的能力。片上集成ONNs是光神经网络微型化发展的典型趋势，它继承了光的优势也保留了其不足。目前所报道的光学神经网络或光电混合神经网络在具体技术上均有所突破，如非线性功能在片上全光神经网络中已经实现[42]，只是现阶段的非线性实现规模还无法支持ONNs在实用化过程中取得较好的表现。因此，无论是全光神经网络还是光电混合神经网络，其发展方向应该是在极大

程度上推动片上集成ONNs的实用化，不断优化关键技术，使ONNs芯片取得更好的表现。目前已有部分研究工作应用到了实际场景中，如Huang等[43]将片上集成ONNs应用于海底光纤链路的非线性补偿，Sludds等[86]基于光子深度学习开发了边缘计算架构Netcast，使得光的自身优势得到了充分发挥，其工作效率在该专用领域超越了电子计算硬件。未来可以将片上集成ONNs的研究方向和目标与实用化场景密切关联，这样即使是专用片上集成ONNs也具有重要的现实应用意义。对于通用片上集成ONNs，需要开发更完整的光学算子和光学算法，这样才能更好地实现ONNs的原位在线训练功能，才能更好地支持ONNs的实用化和通用化。基于此，可以尝试以片上集成ONNs为基本算力单元，进一步构建片上算力网络，从而赋予片上集成ONNs系统更强的算力。从现阶段来看，要实现片上集成ONNs的通用性，并获得类似电子计算硬件的强大功能，片上集成ONNs还需要较长的发展时间。另外，产业界在光学神经网络或光计算方面也开展了研究，如曦智科技、光子算数以及Lightmatter等科技公司相继成立并发布了相关研发产品。其中，曦智科技发布了光子算术计算引擎(PACE)，该PACE的单个光子芯片集成了约10000个光子器件，其运行特定神经网络算法的速度可以超过现阶段电子计算硬件GPU的计算速度[87]。光子算数科技公司在硅光加速计算模块以及光电融合人工智能(AI)加速计算板卡等方面进行了工程化推进。Lightmatter公司发布了Envise、Passage以及Idiom等系列产品[88]，综合考虑了软硬件协同及能耗等方面，有利于将光的固有优势更好地发挥出来。目前产业界的探索尚处于初期阶段，虽然其产品应用还无法像电子计算硬件一样普及，但是其理论研发和工程化相结合的探索过程具有重要意义和价值。关于产业化发展，缩短落地应用周期是非常关键的一环：短周期内可专注于专用场景及领域的工程化推进，以快速实现关键技术及工程化方面的突破；长周期内需要综合考虑软硬件及行业生态链的构建，包括软件、硬件、协议、行业标准、工艺技术等诸多方面。因此，产业界也需要进行多行业互动，如设计、制造、材料、设备以及应用厂商等进行深度合作以推进光计算产品的实用化进程。

4 结束语

近5年来，光学神经网络得到了快速发展，不同的片上集成ONNs计算架构被不断提出，ONNs芯片设计和制造过程中遇到的技术难题也被陆续解决，相信在多交叉学科的共同努力下，片上大规模可重构功能、非线性功能以及片上光功率放大功能等不久后就能实现。目前，虽然片上集成ONNs的实用化方向还未明确，但随着研究的不断深入以及社会需求的不断增加，ONNs的应用场景及实用化方向终会变得清晰。虽然现阶段ONNs不能像电子神经网络一样被广泛应用，但是该阶段ONNs的关键技术积累是有重要意义的。只有不断地深入探索，才能更好地了解ONNs未来在实用化过程中可能遇到的潜在难题，才能够通过目标明确的技术攻关来扫除ONNs在未来实用化过程中遇到的障碍。在推动ONNs发展的过程中，全光神经网络和光电混合神经网络的探索需要并驾齐驱。其中，光电混合神经网络可以结合电辅助的形式暂时绕过光域中难以解决的难题，从而快速优化现有方案，加速光电混合神经网络的实用化过程，尽早寻找到ONNs在实用化过程中的位置和作用，该过程需要考虑的是光电混合ONNs的系统能效，确保ONNs系统带来的优势大于电辅助产生的代价。全光神经网络是ONNs研究的最终目标，因为这样的ONNs才能将光的优势发挥到极致。另外，应该区别对待专用ONNs和通用ONNs。专用ONNs的研究可能仅以应用场景为导向，其目标非常明确，因此研究进展往往会更快。然而，对于通用ONNs的研究，需要综合考虑计算架构、光学算子、光学算法、协议标准、系统软件以及生态构建等，这可能需要较长时间的探索。因此，同时推进专用ONNs和通用ONNs的研究是非常必要的。随着各个学科的不断完善以及交叉学科的深度融合，相信在不同领域研究人员的共同努力和推动下，ONNs在即将到来的人工智能时代定会绽放异彩。

参 考 文 献

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[3] Ananthanarayana T, Srivastava P, Chintha A, et al. Deep learning methods for sign language translation[J]. ACM Transactions on Accessible Computing, 2021, 14(4): 22.

[4] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 26-31, 2013, Vancouver, BC, Canada. New York: IEEE Press, 2013: 6645-6649.

[5] Rausch V, Hansen A, Solowjow E, et al. Learning a deep neural net policy for end-to-end control of autonomous vehicles[C]//2017 American Control Conference (ACC), May 24-26, 2017, Seattle, WA, USA. New York: IEEE Press, 2017: 4914-4919.

[6] Al-Qizwini M, Barjasteh I, Al-Qassab H, et al. Deep learning algorithm for autonomous driving using GoogLeNet[C]//2017 IEEE Intelligent Vehicles Symposium (IV), June 11-14, 2017, Los Angeles, CA, USA. New York: IEEE Press, 2017: 89-96.

[7] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.

[8] Baldi P. Deep learning in biomedical data science[J]. Annual Review of Biomedical Data Science, 2018, 1: 181-205.

[9] Apostolidis K D, Papakostas G A. A survey on adversarial deep

learning robustness in medical image analysis[J]. Electronics, 2021, 10(17): 2132.
[10] Liu Y Q, Qian K, Wang K, et al. Effective scaling of blockchain beyond consensus innovations and Moore's law: challenges and opportunities[J]. IEEE Systems Journal, 2022, 16(1): 1424-1435.
[11] Lin X, Rivenson Y, Yardimci N T, et al. All-optical machine learning using diffractive deep neural networks[J]. Science, 2018, 361(6406): 1004-1008.
[12] Chang J L, Sitzmann V, Dun X, et al. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification[J]. Scientific Reports, 2018, 8: 12324.
[13] Yan T, Wu J M, Zhou T K, et al. Fourier-space diffractive deep neural network[J]. Physical Review Letters, 2019, 123(2): 023901.
[14] Qian C, Lin X, Lin X B, et al. Performing optical logic operations by a diffractive neural network[J]. Light: Science & Applications, 2020, 9: 59.
[15] Miscuglio M, Hu Z B, Li S R, et al. Massively parallel amplitude-only Fourier neural network[J]. Optica, 2020, 7(12): 1812-1819.
[16] Dou H K, Deng Y E, Yan T, et al. Residual D^2NN: training diffractive deep neural networks via learnable light shortcuts[J]. Optics Letters, 2020, 45(10): 2688-2691.
[17] Zhou T K, Lin X, Wu J M, et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit[J]. Nature Photonics, 2021, 15(5): 367-373.
[18] Wang P P, Xiong W J, Huang Z B, et al. Orbital angular momentum mode logical operation using optical diffractive neural network[J]. Photonics Research, 2021, 9(10): 2116-2124.
[19] Mengu D, Zhao Y F, Tabassum A, et al. Diffractive interconnects: all-optical permutation operation using diffractive networks[J]. Nanophotonics, 2023, 12(5): 905-923.
[20] Chen H L, Huang L Z, Liu T R, et al. Fourier Imager Network (FIN): a deep neural network for hologram reconstruction with superior external generalization[J]. Light: Science & Applications, 2022, 11: 254.
[21] Shi W X, Huang Z, Huang H H, et al. LOEN: Lensless opto-electronic neural network empowered machine vision[J]. Light: Science & Applications, 2022, 11: 121.
[22] Li J X, Gan T Y, Bai B J, et al. Massively parallel universal linear transformations using a wavelength-multiplexed diffractive optical network[J]. Advanced Photonics, 2023, 5(1): 016003.
[23] Rahman M S S, Yang X L, Li J X, et al. Universal linear intensity transformations using spatially incoherent diffractive processors[J]. Light: Science & Applications, 2023, 12: 195.
[24] Tait A N, Nahmias M A, Shastri B J, et al. Broadcast and weight: an integrated network for scalable photonic spike processing [J]. Journal of Lightwave Technology, 2014, 32(21): 4029-4041.
[25] Shen Y C, Harris N C, Skirlo S, et al. Deep learning with coherent nanophotonic circuits[J]. Nature Photonics, 2017, 11(7): 441-446.
[26] Tait A N, de Lima T F, Zhou E, et al. Neuromorphic photonic networks using silicon photonic weight banks[J]. Scientific Reports, 2017, 7: 7430.
[27] Hughes T W, Minkov M, Shi Y U, et al. Training of photonic neural networks through in situ backpropagation and gradient measurement[J]. Optica, 2018, 5(7): 864-871.
[28] Bagherian H, Skirlo S, Shen Y C, et al. On-chip optical convolutional neural networks[EB/OL]. (2018-08-09) [2023-05-06]. https://arxiv.org/abs/1808.03303.
[29] Peng H T, Nahmias M A, Lima T F D, et al. Neuromorphic photonic integrated circuits[J]. IEEE Journal of Selected Toptics in Quantum Electronics, 2018, 24(6): 6101715.
[30] Williamson I A D, Hughes T W, Minkov M, et al. Reprogrammable electro-optic nonlinear activation functions for optical neural networks[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 7700412.
[31] Feldmann J, Youngblood N, Wright C D, et al. All-optical spiking neurosynaptic networks with self-learning capabilities[J]. Nature, 2019, 569(7755): 208-214.
[32] Tait A N, Ferreira de Lima T, Nahmias M A, et al. Silicon photonic modulator neuron[J]. Physical Review Applied, 2019, 11 (6): 064043.
[33] Hughes T W, Williamson I A D, Minkov M, et al. Wave physics as an analog recurrent neural network[J]. Science Advances, 2019, 5(12): eaay6946.
[34] Shuhei O, Kasidit T, Shinichi T, et al. Si microring resonator crossbar arrays for deep learning accelerator[J]. Japanese Journal of Applied Physics, 2020, 59(SG): SGGE04.
[35] Zarei S, Marzban M R, Khavasi A. Integrated photonic neural network based on silicon metalines[J]. Optics Express, 2020, 28 (24): 36668-36684.
[36] Qu Y R, Zhu H Z, Shen Y C, et al. Inverse design of an integrated-nanophotonics optical neural network[J]. Science Bulletin, 2020, 65(14): 1177-1183.
[37] Shokraneh F, Geoffroy-Gagnon S, Liboiron-Ladouceur O. The diamond mesh, a phase-error-and loss-tolerant field-programmable MZI-based optical processor for optical neural networks[J]. Optics Express, 2020, 28(16): 23495-23508.
[38] Ong J R, Ooi C C, Ang T Y L, et al. Photonic convolutional neural networks using integrated diffractive optics[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(5): 7702108.
[39] Zhao X M, Lü H B, Chen C, et al. On-chip reconfigurable optical neural networks[EB/OL]. (2021-02-15)[2023-06-05]. https://www.researchsquare.com/article/rs-155560/v1.
[40] Deng H Q, Khajavikhan M. Parity-time symmetric optical neural networks[J]. Optica, 2021, 8(10): 1328-1333.
[41] Zhang H, Gu M, Jiang X D, et al. An optical neural chip for implementing complex-valued neural network[J]. Nature Communications, 2021, 12(1): 457.
[42] Feldmann J, Youngblood N, Karpov M, et al. Parallel convolutional processing using an integrated photonic tensor core [J]. Nature, 2021, 589(7840): 52-58.
[43] Huang C R, Fujisawa S, de Lima T F, et al. A silicon photonic-electronic neural network for fibre nonlinearity compensation[J]. Nature Electronics, 2021, 4(11): 837-844.
[44] Fu T Z, Zang Y B, Huang H H, et al. On-chip photonic diffractive optical neural network based on a spatial domain electromagnetic propagation model[J]. Optics Express, 2021, 29 (20): 31924-31940.
[45] Goi E, Chen X, Zhang Q M, et al. Nanoprinted high-neuron-density optical linear perceptrons performing near-infrared inference on a CMOS chip[J]. Light: Science & Applications, 2021, 10: 40.
[46] Shi Y, Ren J Y, Chen G Y, et al. Nonlinear germanium-silicon photodiode for activation and monitoring in photonic neuromorphic networks[J]. Nature Communications, 2022, 13: 6048.
[47] Zhu H H, Zou J, Zhang H, et al. Space-efficient optical computing with an integrated chip diffractive neural network[J]. Nature Communications, 2022, 13: 1044.
[48] Wang X Y, Xie P, Chen B H, et al. Chip-based high-dimensional optical neural network[J]. Nano-Micro Letters, 2022, 14(1): 221.
[49] Xu S F, Wang J, Yi S C, et al. High-order tensor flow processing using integrated photonic circuits[J]. Nature Communications, 2022, 13: 7970.
[50] Wang Z, Chang L, Wang F F, et al. Integrated photonic metasystem for image classifications at telecommunication wavelength[J]. Nature Communications, 2022, 13: 2131.
[51] Zarei S, Khavasi A. Realization of optical logic gates using on-chip diffractive optical neural networks[J]. Scientific Reports, 2022, 12: 15747.
[52] Yan T, Yang R, Zheng Z Y, et al. All-optical graph representation learning using integrated diffractive photonic computing units[J]. Science Advances, 2022, 8(24): eabn7630.
[53] Ashtiani F, Geers A J, Aflatouni F. An on-chip photonic deep neural network for image classification[J]. Nature, 2022, 606 (7914): 501-506.
[54] Huang Y, Wang W P, Qiao L, et al. Programmable low-threshold

optical nonlinear activation functions for photonic neural networks [J]. Optics Letters, 2022, 47(7): 1810-1813.
[55] Liao K, Li C T, Dai T X, et al. Matrix eigenvalue solver based on reconfigurable photonic neural network[J]. Nanophotonics, 2022, 11(17): 4089-4099.
[56] Luo X H, Hu Y Q, Ou X N, et al. Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible[J]. Light: Science & Applications, 2022, 11: 158.
[57] Ohno S, Tang R, Toprasertpong K, et al. Si microring resonator crossbar array for on-chip inference and training of the optical neural network[J]. ACS Photonics, 2022, 9(8): 2614-2622.
[58] Zhang W P, Huang C R, Peng H T, et al. Silicon microring synapses enable photonic deep learning beyond 9-bit precision[J]. Optica, 2022, 9(5): 579-584.
[59] Bai B W, Yang Q P, Shu H W, et al. Microcomb-based integrated photonic processing unit[J]. Nature Communications, 2023, 14: 66.
[60] Fu T Z, Zang Y B, Huang Y Y, et al. Photonic machine learning with on-chip diffractive optics[J]. Nature Communications, 2023, 14: 70.
[61] Fu T Z, Huang Y Y, Sun R, et al. Integrated diffractive optical neural network with space-time interleaving[J]. Chinese Optics Letters, 2023, 21(9): 091301.
[62] Huang Y, Yue H S, Ma W, et al. Easily scalable photonic tensor core based on tunable units with single internal phase shifters[J]. Laser & Photonics Reviews, 2023, 17(10): 2300001.
[63] Wei M L, Li J Y, Chen Z Q, et al. Electrically programmable phase-change photonic memory for optical neural networks with nanoseconds in situ training capability[J]. Advanced Photonics, 2023, 5(4): 046004.
[64] Huang Y Y, Fu T Z, Huang H H, et al. Sophisticated deep learning with on-chip optical diffractive tensor processing[J]. Photonics Research, 2023, 11(6): 1125-1138.
[65] Liu W C, Fu T Z, Huang Y Y, et al. C-DONN: compact diffractive optical neural network with deep learning regression[J]. Optics Express, 2023, 31(13): 22127-22143.
[66] Poordashtban O, Marzabn M R, Khavasi A. Integrated photonic convolutional neural network based on silicon metalines[J]. IEEE Access, 2023, 11: 61728-61737.
[67] Meng X Y, Zhang G J, Shi N N, et al. Compact optical convolution processing unit based on multimode interference[J]. Nature Communications, 2023, 14: 3000.
[68] Wang J H, Rodrigues S P, Dede E M, et al. Microring-based programmable coherent optical neural networks[J]. Optics Express, 2023, 31(12): 18871-18887.
[69] Pai S, Sun Z H, Hughes T W, et al. Experimentally realized in situ backpropagation for deep learning in photonic neural networks [J]. Science, 2023, 380(6643): 398-404.
[70] Giamougiannis G, Tsakyridis A, Moralis-Pegios M, et al. Neuromorphic silicon photonics with 50 GHz tiled matrix multiplication for deep-learning applications[J]. Advanced Photonics, 2023, 5(1): 016004.
[71] Xu X Y, Tan M X, Corcoran B, et al. 11 TOPS photonic convolutional accelerator for optical neural networks[J]. Nature, 2021, 589(7840): 44-51.
[72] Stelzer F, Röhm A, Vicente R, et al. Deep neural networks using a single neuron: folded-in-time architecture using feedback-modulated delay loops[J]. Nature Communications, 2021, 12: 5164.
[73] Zhang L H, Li C Y, He J Y, et al. Optical machine learning using time-lens deep neural networks[J]. Photonics, 2021, 8(3): 78.
[74] 陈宏伟, 于振明, 张天, 等. 光子神经网络发展与挑战[J]. 中国激光, 2020, 47(5): 0500004.
Chen H W, Yu Z M, Zhang T, et al. Advances and challenges of optical neural networks[J]. Chinese Journal of Lasers, 2020, 47(5): 0500004.
[75] 成骏伟, 江雪怡, 周海龙, 等. 光电智能计算研究进展与挑战[J]. 中国激光, 2022, 49(12): 1219001.
Cheng J W, Jiang X Y, Zhou H L, et al. Advances and challenges of optoelectronic intelligent computing[J]. Chinese Journal of Lasers, 2022, 49(12): 1219001.
[76] 陈蓓, 张肇阳, 戴庭舸, 等. 光学神经网络及其应用[J]. 激光与光电子学进展, 2023, 60(6): 0600001.
Chen B, Zhang Z Y, Dai T G, et al. Photonic neural networks and its applications[J]. Laser & Optoelectronics Progress, 2023, 60(6): 0600001.
[77] Reck M, Zeilinger A, Bernstein H J, et al. Experimental realization of any discrete unitary operator[J]. Physical Review Letters, 1994, 73(1): 58-61.
[78] Clements W R, Humphreys P C, Metcalf B J, et al. Optimal design for universal multiport interferometers[J]. Optica, 2016, 3(12): 1460-1465.
[79] Ribeiro A, Ruocco A, Vanacker L, et al. Demonstration of a 4×4-port universal linear circuit[J]. Optica, 2016, 3(12): 1348-1357.
[80] Tait A N, Wu A X, de Lima T F, et al. Microring weight banks [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2016, 22(6): 312-325.
[81] Tait A N, Jayatilleka H, De Lima T F, et al. Feedback control for microring weight banks[J]. Optics Express, 2018, 26(20): 26422-26443.
[82] Yu N F, Genevet P, Kats M A, et al. Light propagation with phase discontinuities: generalized laws of reflection and refraction [J]. Science, 2011, 334(6054): 333-337.
[83] Zhang J J, Yang J B, Xin H, et al. Ultrashort and efficient adiabatic waveguide taper based on thin flat focusing lenses[J]. Optics Express, 2017, 25(17): 19894-19903.
[84] Wang Z, Li T T, Soman A, et al. On-chip wavefront shaping with dielectric metasurface[J]. Nature Communications, 2019, 10: 3547.
[85] Moughames J, Porte X, Thiel M, et al. Three-dimensional waveguide interconnects for scalable integration of photonic neural networks[J]. Optica, 2020, 7(6): 640-646.
[86] Sludds A, Bandyopadhyay S, Chen Z J, et al. Delocalized photonic deep learning on the internet's edge[J]. Science, 2022, 378(6617): 270-276.
[87] LIGHTELLIGENCE[EB/OL]. [2023-10-16]. https://www.lightelligence.co.
[88] Lightmatter[EB/OL]. [2023-10-16]. https://lightmatter.co.

Review of On-Chip Integrated Optical Neural Networks (Invited)

Fu Tingzhao[1,4,5], Sun Run[2,3], Huang Yuyao[2,3], Zhang Jianfa[1,4,5], Yang Sigang[2,3], Zhu Zhihong[1,4,5], Chen Hongwei[2,3*]

[1]*College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha* 410073, *Hunan, China*;

[2]*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*;

[3]*Beijing National Research Center for Information Science and Technology, Beijing* 100084, *China*;

[4]*Hunan Provincial Key Laboratory of Novel Nano-Optoelectronic Information Materials and Devices, National University of Defense Technology, Changsha* 410073, *Hunan, China*;

[5]*Nanhu Laser Laboratory, National University of Defense Technology, Changsha* 410073, *Hunan, China*

Abstract

Significance With the advent of the era of artificial intelligence, advanced algorithms represented by deep learning algorithms are rapidly developing, driven by big-data resources. This is promoting the extensive application of neural networks in various fields of social development, including computer vision, natural language processing, speech recognition, automatic driving, and biomedicine. In the past two decades, advanced semiconductor technology has led to the creation of various types of computer hardware with excellent performances, which meet the computing capacity resource requirements of neural networks in various fields.

However, with the continuous elevation of social intelligence in the future, neural networks will require even greater computing resources when processing complex tasks. Simultaneously, the machining accuracy of semiconductor process technology has approached the physical limit, and ultra-small on-chip devices are susceptible to quantum tunneling and thermal effects, which may prevent the proper operation of chips manufactured with this machining accuracy. Hence, it will be difficult to continue to increase computing capacity resources by further improving the processing accuracy of semiconductor processes. Consequently, it is imperative to find a new computing paradigm to replace the existing computing architecture to break through this computing-capacity bottleneck.

An optical neural network (ONN) is a high-performance novel computing paradigm that differs from von Neumann computing schemes. It has advantages such as low latency, low power consumption, large bandwidth, and parallel signal processing. Its inference process relies on the diffraction and interference of light, and no additional energy supply is required for the entire calculation process. Compared with traditional electronic hardware, it has natural advantages in performing large-scale linear matrix operations.

Progress This study comprehensively reviews the research progress and challenges related to on-chip integrated ONNs. These are typically designed based on a Mach-Zehnder interferometer (MZI), micro-ring resonator (MRR), or subwavelength unit (SWU). When first introduced, the on-chip ONNs are based on MZIs (Figs. 1 and 2), which can achieve matrix operations in the inference process by combining the topological cascading and matrix decomposition methods of MZIs. Next, on-chip ONNs based on MRRs are presented (Figs. 3 and 4). MRRs can redistribute the optical power at different frequencies, and the matrix operation function in the ONN inference process can be actualized by cleverly designing the weights at different wavelengths after filtering. Then, on-chip diffractive optical neural networks (DONNs) based on SWUs are introduced (Figs. 5 and 6). This kind of ONN can realize the wavefront modulation of the propagating light in the slab waveguide by designing the sizes of the SWUs to obtain specific diffraction results to complete reasoning tasks. Finally, we compare the integration, energy consumption, and computational throughput of on-chip ONNs designed with different structural units based on experiments with integrated ONN chips (Table 1). The above research provides a valuable reference for the exploration of on-chip ONNs.

Conclusions and Prospects On-chip ONNs designed based on MZIs or MRRs both have reconfigurable functions, and these basic structural units, MZIs and MRRs, can be further combined with phase change material (PCM) units to achieve nonlinear functions on the ONN chips. However, the matrix scale that these ONNs can handle in parallel is often relatively low. In contrast, an on-chip DONN designed based on SWUs can process large-scale matrices in parallel because of its small size, high integration, and easy large-scale expansion. Nevertheless, it is eminently challenging to implement reconfigurable and nonlinear functions on a DONN chip. Therefore, achieving reconfigurable functions, nonlinear functions, and the parallel processing of large-scale matrices on ONN chips requires joint efforts from multiple disciplines. In the future, the development direction of on-chip ONNs is supposed to be closely related to their practical applications. Meanwhile, it will be better to promote the research of both dedicated on-chip ONNs and general on-chip ONNs. Dedicated on-chip ONNs are designed for specific application scenarios, which may rapidly propel the research progress. Universal on-chip ONNs require an inclusive consideration of the computing architecture, optical operators, optical algorithms, protocol standards, system software, and ecological construction, with the goal of laying a solid foundation for the generalization of ONN chips. With continuous improvements in various disciplines and the deep collaboration of interdisciplinary fields, on-chip ONNs will shine brightly in the upcoming era of artificial intelligence through the joint efforts of researchers in all trades and professions.

Key words integrated optics; optical computing; optical neural networks; chip; artificial intelligence