激光与光电子学进展 先进成像

光学神经网络及其应用

陈蓓[1]，张肇阳[1]，戴庭舸[2]，余辉[1,3]，王曰海[1]，杨建义[1*]

[1]浙江大学信息与电子工程学院，浙江 杭州 310027；

[2]浙江大学宁波理工学院，浙江 宁波 315100；

[3]之江实验室，浙江 杭州 310027

摘要 由于光传输具备高通量、低延迟、低能耗等优势，光学神经网络有望应对目前人工智能技术发展中所面临的能耗和计算效率的挑战，成为近年来学术界和工业界的研究热点。光学神经网络的目标在于用光子作为物理载体构建人工神经网络算法中的基本计算单元，从而实现高性能的新型计算架构，并将其应用于实际问题的解决。本综述介绍了光学神经网络中关键光子器件的工作原理和特点、系统架构特征与应用场景。在跟踪大量国内外研究进展后，进一步分析了光学神经网在系统实现上所面临的挑战及发展趋势。

关键词 光计算；光学神经网络；线性矩阵计算；非线性激活器

中图分类号 O436 **文献标志码** A **DOI**：10.3788/LOP222304

Photonic Neural Networks and Its Applications

Chen Bei[1], Zhang Zhaoyang[1], Dai Tingge[2], Yu Hui[1,3], Wang Yuehai[1], Yang Jianyi[1*]

[1]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou* 310027, *Zhejiang, China*;

[2]*Ningbo Research Institute, Zhejiang University, Ningbo* 315100, *Zhejiang, China*;

[3]*Zhejiang Lab, Hangzhou* 310027, *Zhejiang, China*

Abstract Photonic neural networks (PNNs) are proposed to balance the demands between a substantial increase in computation ability and a decrease in computing power consumption, owing to the superiorities in terms of large bandwidth, low latency, and low power consumption in optical transmission. Hence, in recent years, PNNs have become the research hotspot both in academia and industry. By utilizing photons as the physical media, the basic computing units in artificial neural network algorithms can be built and experimentally demonstrated. In further, PNNs might be adopted as a new computing architecture with high performances and be applied to solve the practical applications. In this paper, the working principle and characteristics of the core optical devices in PNNs are described, along with the system architecture and application scenario. In addition, the present challenges and future development trends of PNNs are discussed, after reviewing the research progress of PNNs at home and aboard.

Key words optical computing; photonic neural network; linear matrix multiplication; nonlinear activator

1 引 言

当今世界处于一个信息爆炸的时代，如何像人脑一样快速高效地实现信息处理，成为了信息技术领域的重要研究课题。20世纪50年代，人工智能(AI)概念[1]被正式提出。在迄今半个多世纪的发展历程中，以人工神经网络(ANN)为核心的AI技术蓬勃发展，成为学术界和工业界的研究与应用焦点。其中，海量数据的获取与存储、不同神经网络(NN)算法架构的实现、高速低功耗的高性能计算等都依赖于AI技术的物理载体——芯片。

阿西莫夫(Asimov)人工智能研究所根据不同的网络连接方式汇总了当前典型的ANN算法模型[2]，如图1(a)所示。目前，ANN算法主要运行在以电子为物理载体的硬件架构中，包括基于冯·诺依曼架构和串行逻辑处理的中央处理器(CPU)、基于并行处理的图形

收稿日期：2022-08-15；**修回日期**：2022-09-15；**录用日期**：2022-09-23；**网络首发日期**：2022-10-08

基金项目：国家自然科学基金(61775196，62005242)、国家重点研发计划(2021YFB2801801)

通信作者：*yangjy@zju.edu.cn

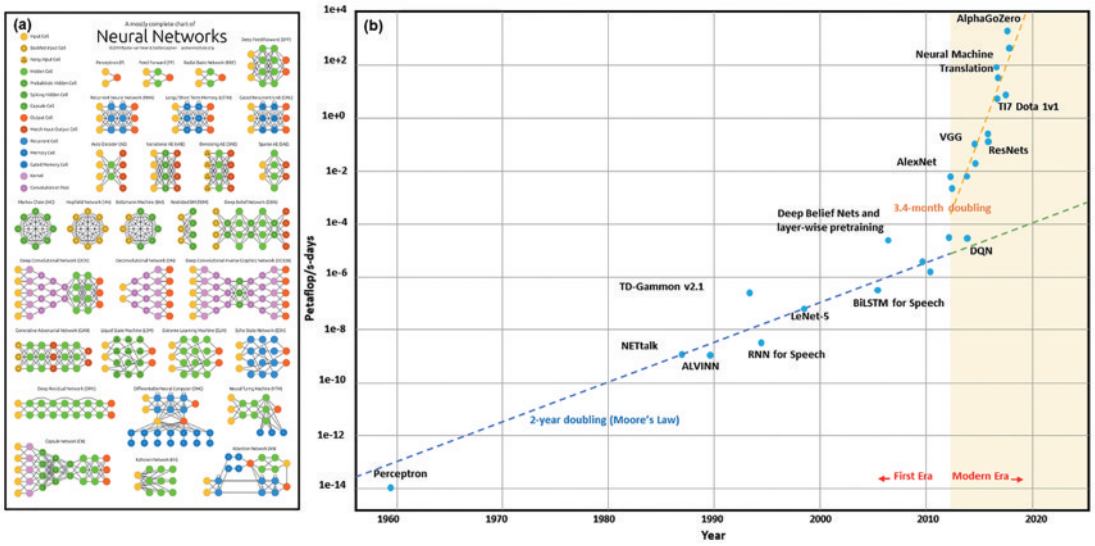

图 1　人工智能技术的发展。(a)典型的 ANN 算法模型[2];(b)AI 算力需求与摩尔定律间的对比[10]

Fig. 1　The development of artificial intelligence. (a) Typical architectures of ANN[2]; (b) comparison between total amount of AI compute and Moore's law[10]

处理器(GPU)、现场可编程门阵列(FPGA),例如:百度公司推出的昆仑芯片[3]、专用集成芯片(ASIC);谷歌公司推出的张量处理器(TPU)[4]和中国科学院计算技术研究所推出的寒武纪系列加速芯片[5-6]、类脑计算芯片;IBM 公司推出的 TrueNorth 芯片[7]和清华大学团队研发的"天机"芯片[8];等等。根据硅谷人工智能研究组织 OpenAI 统计,AI 算力需求的增长可以划分为两个阶段[9-10],如图 1(b)所示。从 1959 年—2012 年,AI 算力需求的增长和摩尔定律增速相当,意味着电子芯片可以满足 AI 算力增长的需要。从 2012 年至今,AI 计算中的浮点计算量以指数速率快速增长,算力需求每 3.4 个月增长一倍,远远超越摩尔定律[11]每 18~24 个月增长一倍的速度,同时也伴随着能耗问题。庞大的算力需求给现有的电子处理器带来极大的压力,亟待研究新模式的硬件架构来解决摩尔定律增速与算力需求增速之间的矛盾。

得益于光学传输具备大带宽、低损耗、低延时、低功耗、可高度并行的优势,光计算有望在多方面解决目前 AI 技术发展中所面临的能耗和计算效率的问题,成为未来 AI 计算硬件架构的主流实现方式之一。

2　光计算及光学神经网络简介

广义的光计算(optical computing)包括所有利用光子作为物理载体来实现计算的架构方案,通常可分为数字光计算(digital optical computing)和模拟光计算(analog optical computing),其主要实现方式包括空间光学和集成光学两大类,如图 2 所示。

其中,关于数字光计算的研究可以追溯到 20 世纪 50 年代[12]。当时,数字电子计算机取得巨大成功,极大激发了科研工作者对于数字光子计算机的研究热情。其主要思路是基于布尔逻辑[13],设计光子晶体管[14]来代替电子晶体管,从而实现光学逻辑门运算[14-20]。受限于光学逻辑电平的恢复与存储、光学器件集成密度低等问题[21],基于光学逻辑门的数字光计算并没有被验证为是一种实际有效的计算架构。

近年来,基于光自身物理特性的模拟光计算架构不断被提出,其直接利用光在自由空间或某种介质传输过程中产生的幅度、相位等光学特性变化来实现特定功能的计算。根据计算功能的类型,模拟光计算可以分为光学可编程信号处理器(photonic programmable signal processors)[22-29]、光学伊辛机(photonic Ising computing)[30-44]、光学神经网络(photonic neural networks)[45-95]等 3 大类。其中:光学可编程信号处理器可用于实现微波光子信号处理、光学带宽可调滤波器、任意波形发生等各种自定义功能模块;光学伊辛机主要用于高效求解如电路设计、路径优化等 NP-Hard[96]组合优化问题;光学神经网络可以构建不同 ANN 算法架构中的线性矩阵计算、非线性函数激活等基本算子,从而实现光脉冲神经网络(photonic spiking neural networks)[45-54]、光学循环神经网络(photonic recurrent neural networks)[55-72]、光学多层感知机(photonic multilayer perceptrons)[73-82]以及光学卷积神经网络(photonic convolutional neural networks)[54,74,81,83-95]。其中,光学储备池计算(photonic reservoir computing)[55-72]是光学循环神经网络中的重要分支。

如上文所述,光学神经网络作为光计算领域的重要研究方向之一,同时也是 AI 计算中亟待探索的新型硬件架构之一。本综述将主要关注光计算领域中光学神经网络这一分支,跟进其中光学多层感知机和光学卷积神经网络的研究进展。图 3 梳理了在人工神经网络发展过程中光学神经网络的演变,并标注了人工神经网络和光学神经网络两者在其发展过程中的一些重

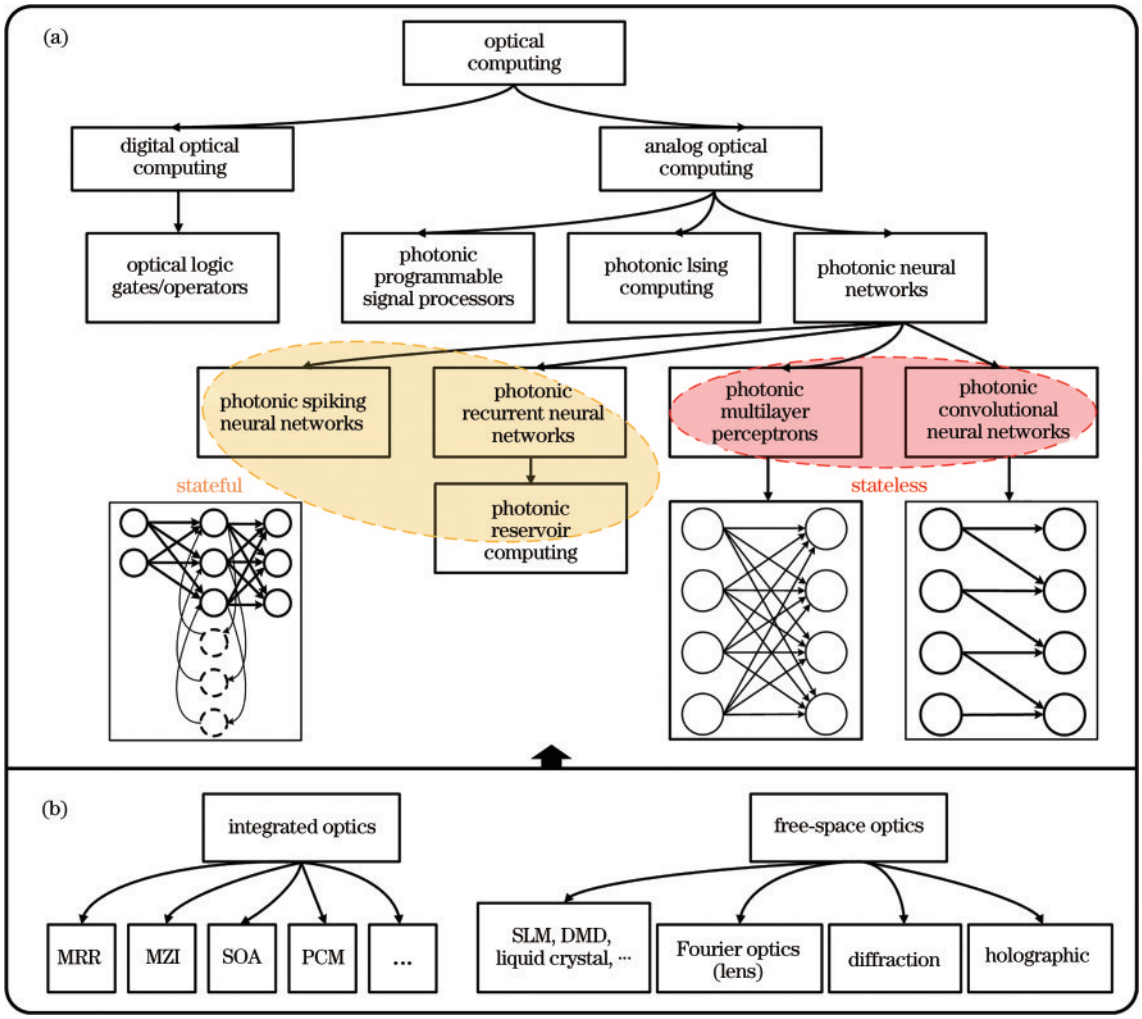

图 2　光计算。(a)光计算分类;(b)光计算的主要实现方式

Fig. 2　Optical computing. (a) Classification of optical computing; (b) implementations of optical computing

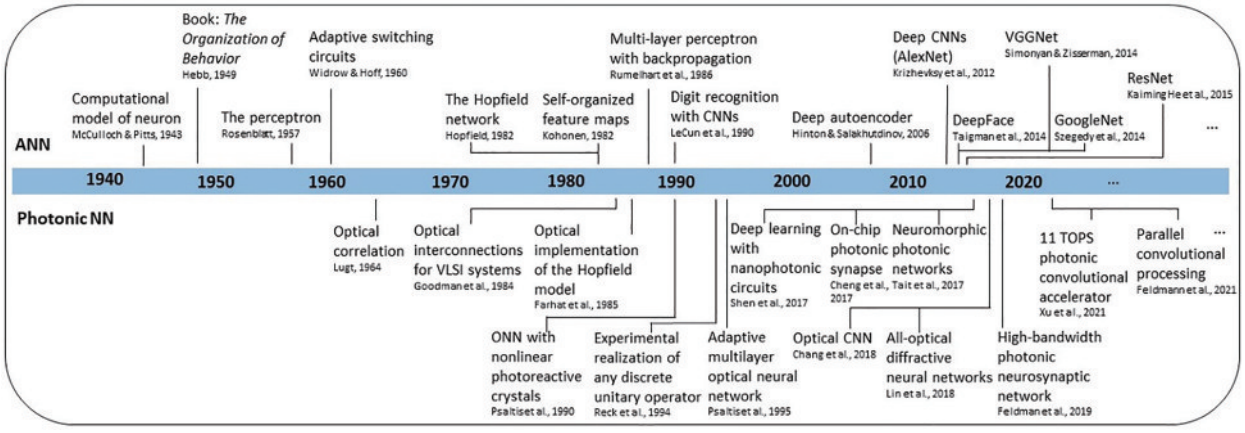

图 3　人工神经网络发展过程[97-112]中光学神经网络的演变[51,73-74,85,93-94,113-123]

Fig. 3　Evolution of PNNs[51,73-74,85,93-94,113-123] during the development of ANNs[97-112]

要突破节点。1943 年,美国心理学家 McCulloch 等[97]从信息处理的角度研究神经细胞行为的数学模型表达,提出了二值神经元模型,并命名为 MP 模型。MP 模型的提出开启了人们对于神经网络的研究。1949 年,心理学家 Hebb[98]提出著名的 Hebb 学习规则,即由神经元之间结合强度的变化来实现神经元学习的方法。1957 年,Rosenblatt[99]提出感知机模型(perceptron),从工程角度出发,研究了用于信息处理的神经网络模型,其基本符合神经生理学的原理。1960 年,Widrow 等[100]提出了自适应线性元件,该元件是一个连续取值的线性阈值网络,在信号处理系统中应用十分广泛。在大量的早期开创性工作[97-102]之后,Rumelhart 等[103]

在多层神经网络模型的基础上，提出了多层神经网络模型中权重的反向传播(BP)学习算法，解决了多层神经网络的学习问题，实验证明BP算法具有很强的学习能力，可以完成不同的学习任务，从而解决许多实际问题。然而，在20世纪90年代—2006年，人工神经网络的研究与发展经历了一段沉寂期，在此期间深度学习三剑客Hinton、LeCun和Bengio[104-106]依然坚持着对该领域的研究。直至2006年，Hinton等[107]提出一种自动编码机(auto-encoder)的方法，部分解决了神经网络参数初始化的问题。自此，2006年被视为深度学习的起始元年，发展至今深度学习也一直受到学术界和工业界的极大关注，层出不穷的人工神经网络算法和框架被提出[108-112]。值得关注的是，在人工神经网络的演变过程中，光学神经网络也伴随着光学卷积[113]、光学线性矩阵乘法[114-115]、光学非线性函数[116]等基本算子的提出和光互连[117]的广泛应用得到发展。1985年，宾夕法尼亚大学Farhat和加州理工学院的Peak课题组[118]联合发表了第一个实验实现的光学神经网络。该光学神经网络实现了包含32个神经元且具有反馈机制的全连接网络，这一进展激发了学术界关于光学神经网络的研究热情。20世纪90年代，光学神经网络架构的实现方式主要集中在空间光学上[119-121]，其充分利用了空间光的并行性。2017年，麻省理工学院Shen等[73]发表了基于56个可编程马赫-曾德尔干涉仪(MZI)拓扑级联架构的片上集成光学神经网络。该架构实现了矩阵-向量乘法的线性运算，并实验演示了2层全连接神经网络对于4个元音信号的识别实验，硬件识别准确率达76.7%。同年，普林斯顿大学Tait等[122]发表了基于并联级联的微环谐振器(MRR)结构的片上硅基光循环神经网络。同期，牛津大学Cheng等[123]发表了基于相变材料(PCM)的权重可调且非易失的片上光神经元。这3项研究均为基于光子集成回路实现神经网络中推理任务的光学方案，为实现高性能、高集成度、低功耗的光神经网络提供了可行性。随后，各种新型的光学神经网络架构及其应用[51,74,85,93-94]得到井喷式的发展。

3　光学神经网络中的关键光子器件及其系统架构与应用

不同架构的ANN算法由多种基本计算单元通过多样的连接方式组合而成[2-124]，其中，最常用的基本算子有矩阵-向量乘的线性矩阵计算、卷积计算、非线性激活函数、池化、等等。光学神经网络的目标在于：用光子作为物理载体构建ANN算法中的基本计算单元，充分发挥其高速、低功耗、低延时、高通量等特点，从而实现高性能的AI计算架构，并将其应用于实际问题的解决。本综述从面向光学神经网络的关键光子器件、系统架构与应用两方面进行梳理，跟踪国内外研究进展，分析光学神经网系统实现中所面临的挑战及其发展趋势。其中：关键光子器件的梳理主要集中于矩阵-向量乘的线性矩阵计算与非线性激活器两部分；在系统架构与应用中，主要介绍已报道的应用领域，同时也梳理了光学神经网络系统的完整设计流程，其自上而下包括目标应用确定、数据集标定、网络参数训练与性能评估、硬件参数确定与性能评估、硬件架构设计、基本光学算子实现等6部分。

3.1　光学神经网络中线性矩阵计算的研究现状和发展趋势

线性矩阵计算作为ANN的基本算子，占据了大部分的计算任务。例如对于GoogleNet[111]和OverFeat[125]模型来说，线性矩阵计算的计算量超过了80%[126]。因而，提高线性矩阵计算的性能是新型AI计算硬件架构的必要需求之一。相较于电子架构，光学线性矩阵计算在计算速度、信号延时、计算密度和功耗等方面展现了更大的优势。

光学线性矩阵计算主要是为了实现矩阵-向量乘法，即$\boldsymbol{Y}=\boldsymbol{W}\times\boldsymbol{X}$，如图4所示。其中，$\boldsymbol{W}$为权重矩阵，$\boldsymbol{X}$为输入向量，$\boldsymbol{Y}$为输出向量。按照其工作原理和实现方式，主要可以分为3类：1)基于空间光结构的光学线性矩阵计算；2)基于片上相干原理的光学线性矩阵计算；3)基于波分复用(WDM)技术的光学线性矩阵计算。

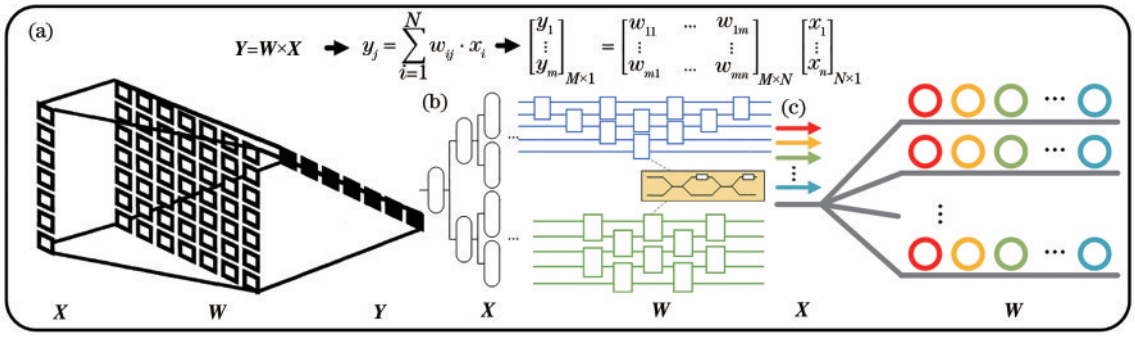

图4　光学线性矩阵计算。(a)空间光结构；(b)片上相干原理；(c) WDM技术

Fig. 4　Optical linear matrix multiplication. (a) Free-space optics; (b) integrated coherent optics; (c) WDM optics

3.1.1　基于空间光结构的光学线性矩阵计算

基于空间光结构的光学线性矩阵计算主要基于空间光调制器(SLM)[76,127-128]、数字微镜器件(DMD)[85,129]、光学衍射板[74,80,130]等空间光学器件实现，如图5所示。

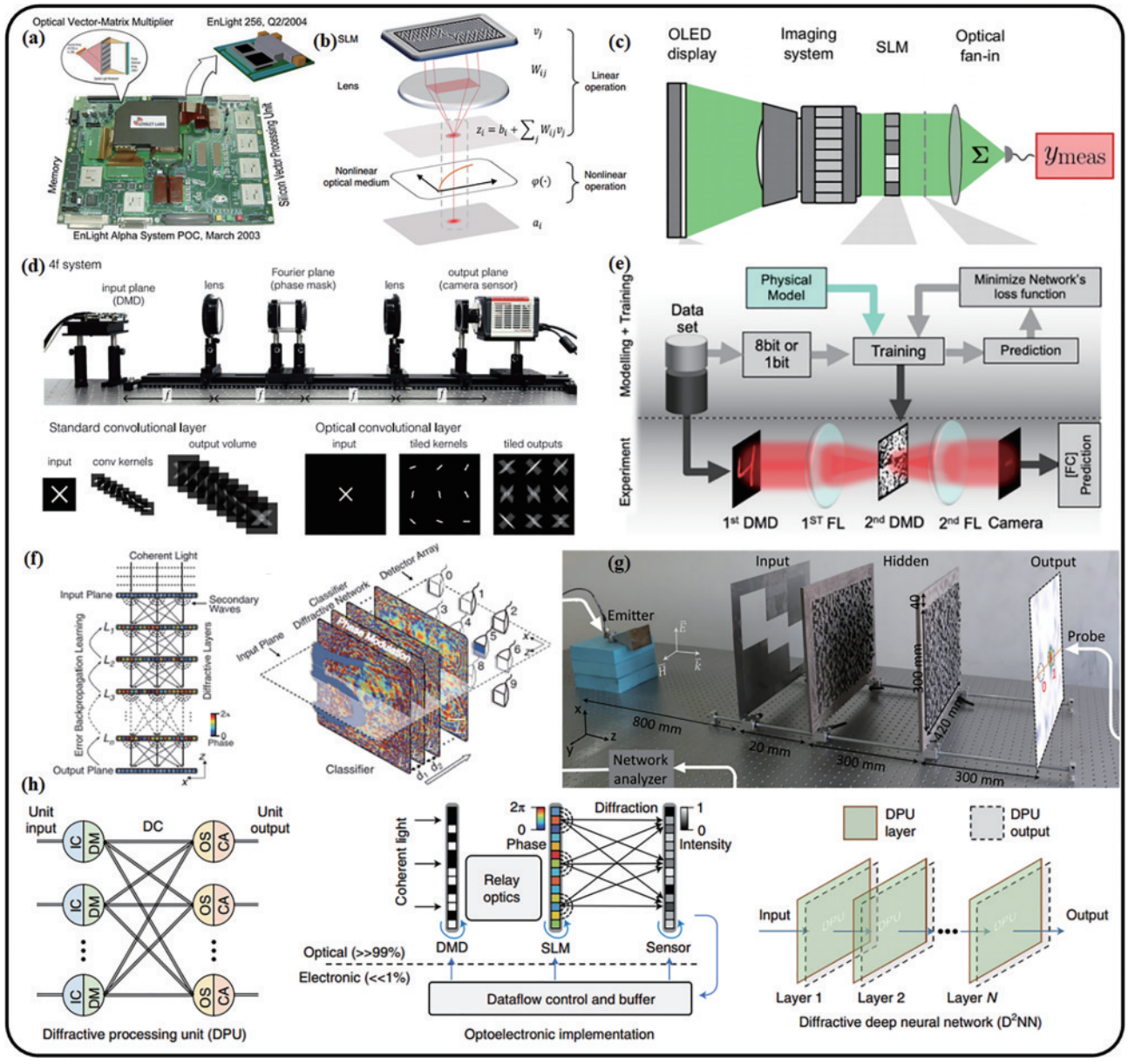

图 5　基于空间光结构的光学线性矩阵计算。(a)～(c)SLM[76,127-128]；(d)(e) DMD[85,129]；(f)～(h)光学衍射板[74,80,130]

Fig. 5　Optical linear matrix multiplications based on free-space optics. (a)–(c) SLM[76,127-128]; (d) (e) DMD[85,129]; (f)–(h) diffractive optics[74,80,130]

3 种具有代表性的基于 SLM 的光学线性矩阵计算如图 5(a)～(c)所示。2004 年，以色列 Lenslet 公司推出了首款光学数字信号处理器(ODSP)——EnLightTM[256][127]，如图 5(a)所示，其展示了超快速数字信号处理器能力，每秒可以执行 8 万亿次的乘积累加运算(MAC)操作，处理速度为普通数字信号处理器的数千倍。EnLightTM[256] 主要包括 3 个核心组件：能执行超高速向量矩阵操作的光学矢量矩阵乘法器(VMM)、每秒能执行 1280 亿次操作的向量处理器(VPU)以及用于系统控制和其他处理的 DSP。其中，VMM 属于光域组件，VPU 和 DSP 属于电域组件。2019 年，香港科技大学杜胜望课题组[76]发表了基于 SLM 傅里叶光学变换的矩阵-向量乘法，并由此构建了两层全连接的全光神经网络，如图 5(b)所示。其中，权重矩阵元素 w_{ij} 可以在 10 次迭代周期内达到高于 95% 的配置准确率，且输入向量元素 x_{ij} 的配置均方误差小于 0.017。2022 年，康奈尔大学 Wang 等[128]提出了一种功耗小于一个光子且手写数字识别准确率达 99% 的光学神经网络架构，如图 5(c)所示。该架构中的矩阵-向量乘法由有机发光二极管(OLED)阵列和 SLM 组合实现，可并行实现 505521(711 × 711)次线性乘法计算。

此外，图 5(d)、(e)展示了 2 种由 DMD 构建的光学 4f 成像系统实现的光学线性矩阵计算架构。图 5(d)是斯坦福大学 Wetzstein 课题组[85]于 2018 年发表的基于计算成像架构的光电混合卷积神经网络。该方案基

于线性的、空间不变的4*f*成像系统实现光学卷积(optical correlator)计算，其能耗趋于0。同时，可以通过训练光学掩膜版上的相位实现权重的优化。2020年，乔治华盛顿大学Sorger课题组[129]发表了基于4*f*光学系统的空间滤波原理实现的幅度可调光学卷积神经网络架构，如图5(e)所示。该架构中的4*f*成像系统使用高分辨率的DMD和傅里叶透镜对实现，可以达到10^6次的通道并行度，在手写数字识别的应用上达到98%的准确率。

图5(f)～(h)是3种具有代表性的基于光学衍射板的线性矩阵计算方案。其中，图5(f)是加州大学洛杉矶分校Ozcan课题组[74]于2018年发表的利用3D打印技术实现的全光衍射深度神经网络(D^2NN)。该方案在手写数字识别这一应用上的准确率可达到91.75%。权重矩阵$\boldsymbol{W}$配置在由3D打印技术制造的光学衍射板上，光在透过衍射板时会由于衍射板不同位置的透射或反射系数不同导致光的相位和幅度变化。该配置方法在数值上遵循Rayleigh-Sommerfeld diffraction equation[131]，可实现神经网络的推理过程，其能耗接近于0。2020年，浙江大学陈红胜课题组[80]发表了一种可用于光学逻辑门实现的全光衍射神经网络，如图5(g)所示。同样基于Rayleigh-Sommerfeld diffraction equation，该架构利用复合的惠更斯超表面结构[132]来实现波前的幅度和相位控制，从而完成全连接网络中的权重配置，并演示了光学非门、或门和与门的逻辑运算实验。2021年，清华大学戴琼海课题组[130]发表了大规模可重构的光电混合衍射处理器，如图5(h)所示。利用DMD模块量化输入数据并通过电光转换为复值光场，在SLM上通过调制光场的幅度或相位实现编码。不同的输入神经元通过光学衍射连接到各个输出神经元，同时控制连接强度的突触权重由波前衍射调制决定。

3.1.2　基于片上相干原理的光学线性矩阵计算

利用光学相干性在光子集成回路中实现的光学线性矩阵计算是一种紧凑的光学矩阵-向量乘法实现方法，其主要结构形态包含了可编程MZI拓扑级联[73,82-83,133]、可配置的推挽式幅度调制器[134]以及片上衍射单元和可编程MZI的组合[95]，如图6所示。

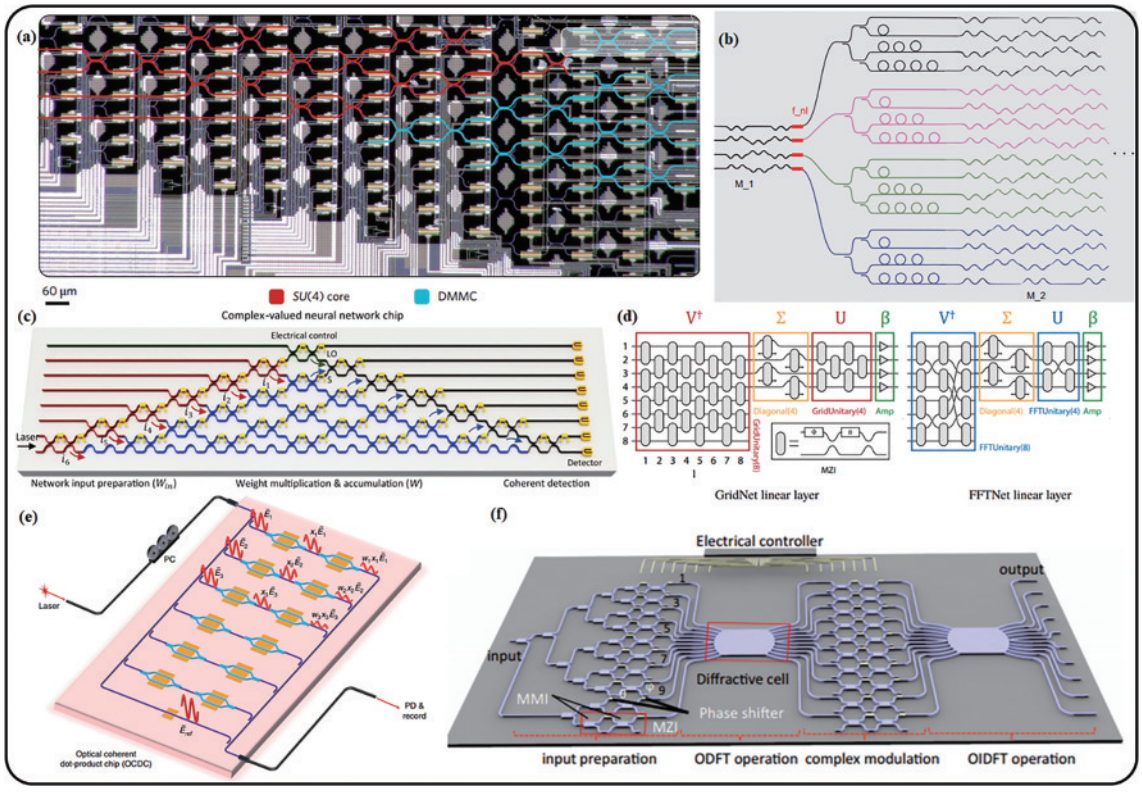

图6　基于片上相干原理的光学线性矩阵计算。(a)～(d)可编程MZI拓扑级联[73,82-83,133]；(e)可配置的推挽式幅度调制器[134]；(f)片上衍射单元和可编程MZI组合[95]

Fig. 6　Optical linear matrix multiplications based on integrated coherent optics. (a)–(d) Programmable MZI arrays[73,82-83,133]; (e) configurable push-pull modulators[134]; (f) combination of on-chip diffractive cell and programmable MZI[95]

对于可编程MZI拓扑级联的相干光架构来说，由于其无法直接加载任意规模的权重矩阵$\boldsymbol{W}$，需要先将权重矩阵$\boldsymbol{W}$进行矩阵奇异值分解(SVD)[135]，将其拆解为两个酉矩阵$\boldsymbol{U}$、$\boldsymbol{V}$和对角矩阵$\boldsymbol{\Sigma}$，即$\boldsymbol{W}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$，继而通过可编程的MZI拓扑级联的方式实现任意规模的酉矩阵$\boldsymbol{U}$、$\boldsymbol{V}$，并采用相位可调波导实现对角矩阵$\boldsymbol{\Sigma}$，进而实现权重矩阵$\boldsymbol{W}$的配置。2017年，麻省理工学院Shen等[73]发表了基于MZI拓扑级联的矩阵-向量乘法

器，利用相干光和矩阵SVD分解的方式，实现了片上集成的全连接光学神经网络架构，如图6(a)所示。由于光检测器噪声、热串扰、MZI级联过程中的误差累积等因素，其在4个元音的语音识别问题上硬件准确率仅为76.7%。2018年，该课题组进一步发表了基于MRR延时与MZI拓扑级联结合的光学卷积神经网络[83]，如图6(b)所示。该架构通过MRR产生的延时重排卷积核，从而将矩阵卷积计算转换为矩阵-向量乘法，多路并行的MRR延时和MZI拓扑级联组合可实现多个卷积核的并行计算。2021年，南洋理工大学Liu课题组[82]在同样的MZI拓扑级联架构中增加了相位调控这一维度，基于可编程幅度和相位的MZI结构实现了全连接的复数光学神经网络，如图6(c)所示。该架构将手写数字识别的准确率提升至90.5%。另外，考虑到工艺误差对于大规模的MZI拓扑级联的影响，2019年加州伯克利大学和英特尔公司联合提出误差鲁棒性较好的光学神经网络架构[133]，如图6(d)所示。其中，线性矩阵计算可以由调节度更高的GridNet或容错性更好的FFTNet架构实现。该工作同时搭建了算法仿真系统，可以在制作工艺之前对器件进行误差分析和优化，对大规模光学神经网络的流片制备和系统误差评估起到指导作用。综上所述，图6(a)～(d)均为依赖矩阵SVD分解的MZI拓扑级联架构，其权重矩阵$\boldsymbol{W}$中的任一元素w_{ij}无法直接对应于单个光器件，即其无法由单器件直接独立控制和实现，而是需要通过多个MZI拓扑级联的方式来构建其在经过矩阵SVD分解后的中间变换矩阵(酉矩阵$\boldsymbol{U}$、$\boldsymbol{V}$和对角矩阵$\boldsymbol{\Sigma}$)。其中，酉矩阵的构建方式可由不同的分解算法[114-115,133]来实现，对应到物理实现上为不同的MZI拓扑连接方式。

为了简化上述架构的复杂度，同时增加架构在可重构上的灵活性，上海交通大学邹卫文课题组[134]于2021年发表了基于推挽式幅度调制器实现的光学相干点乘计算芯片，如图6(e)所示。该架构通过幅度可调的推挽式调制器可实现片上矩阵-向量乘法和卷积计算两种计算方式，其数值表征范围可以从非负数域拓展到完整的实数域上，且权重矩阵中的任一元素w_{ij}可由单个调制器独立控制。另外，为实现更紧凑的相干光线性矩阵计算，南洋理工大学Liu课题组[95]于2022年发表了基于片上集成光学衍射单元和可编程MZI组合的架构方案，该方案能实现并行傅里叶变换和卷积计算，如图6(f)所示。在该方案中，原需N^2个MZI级联完成的线性矩阵计算，可以由2个基于平面波导的超紧凑衍射单元和N个MZI实现，极大程度地缩小了芯片尺寸，并降低了计算功耗。

3.1.3　基于WDM技术的光学线性矩阵计算

基于WDM技术的光学线性矩阵计算是一种无需矩阵分解的、可直接通过波长和矩阵元素一一对应的、同时大规模可拓展的并行光学矩阵-向量乘法实现方式。输入向量$\boldsymbol{X}$中的元素x_i对应一个具有特定频率或波长的输入光，然后通过$m\times n$规模的光学权重矩阵阵列对其进行不同的幅度调节，从而实现对输入信号的加权，即$y_j=\sum_{i=1}^{n}w_{ji}\cdot x_i$。基于WDM技术的光学线性矩阵计算的物理实现方式包括基于并联级联的MRR[46,89,122]、PCM[51,81,92]、半导体光放大器(SOA)[136]、色散光纤[137]以及光频梳[93-94]，如图7所示。

图7(a)、(b)为基于并联级联MRR的光学线性矩阵计算架构。2014年，普林斯顿大学Tait等[46]提出了基于并联级联MRR的神经元权重并行配置架构，如图7(a)所示。该架构被命名为广播-权重方法(broadcast and weight)，其中，每个输入元素可以被独立调制为任意实数。另外，该课题组[122]于2017年发表了这一架构的实验测试结果，并演示了光学循环神经网络的实现。2020年，该课题组[89]通过对输入图像进行分块处理以及对卷积核进行矩阵-向量转换和重排，在这一架构上实现了光学卷积操作，如图7(b)所示。

图7(c)～(e)展示了3种典型的基于PCM的光学线性矩阵计算架构。2018年，乔治华盛顿大学Miscuglio等[81]发表了基于PCM的光子张量核，如图7(c)所示。输入向量由高速调制器编码，张量核通过在串联双环的中间耦合波导上集成非易失的PCM来实现，进而由光电探测器(PD)进行非相干求和。其中，串联双环用以实现不同波长的选择，非易失PCM用以进行权值控制和保持。2019年，德国明斯特大学Pernice课题组[51]发表了基于PCM和MRR的全光神经网络，如图7(d)所示。其中，单路PCM和MRR实现了权重矩阵中每个元素的配置，复用总线上的PCM和MRR实现了加权求和后的全光非线性激活。2021年，华盛顿大学李墨课题组[92]发表了基于PCM的多模光学卷积神经网络，如图7(e)所示。输入向量由可变光衰减器(VOA)实现编码，卷积核在多模氮化硅(SiN)波导上集成PCM进行模式转换从而实现权值配置，最终可实现6位比特精度的实数编码。

另外，荷兰埃因霍温大学Stabile课题组[136]于2020年实现了基于磷化铟(InP)SOA的多路并行全连接光学神经网络，如图7(f)所示。该架构中光学线性矩阵计算的实现方式为：通过阵列波导光栅(AWG)对总线上复用的多波长进行单一波长分离，进而由每一个SOA实现各个波长的权重配置，最后由PD进行叠加求和。同年，清华大学陈宏伟课题组[138]基于对超短脉冲的时域拉伸实现权重矩阵和输入向量的配置，从而实现全连接的光学神经网络[137]，如图7(g)所示。该架构的核心是利用光纤色散特性实现线性矩阵运算，通过并行变串行的方案实现了光电混合的全连接神经网络。

图7(h)、(i)是两种典型的基于光频梳的光学线性矩阵计算架构。其中，图7(h)是德国明斯特大学

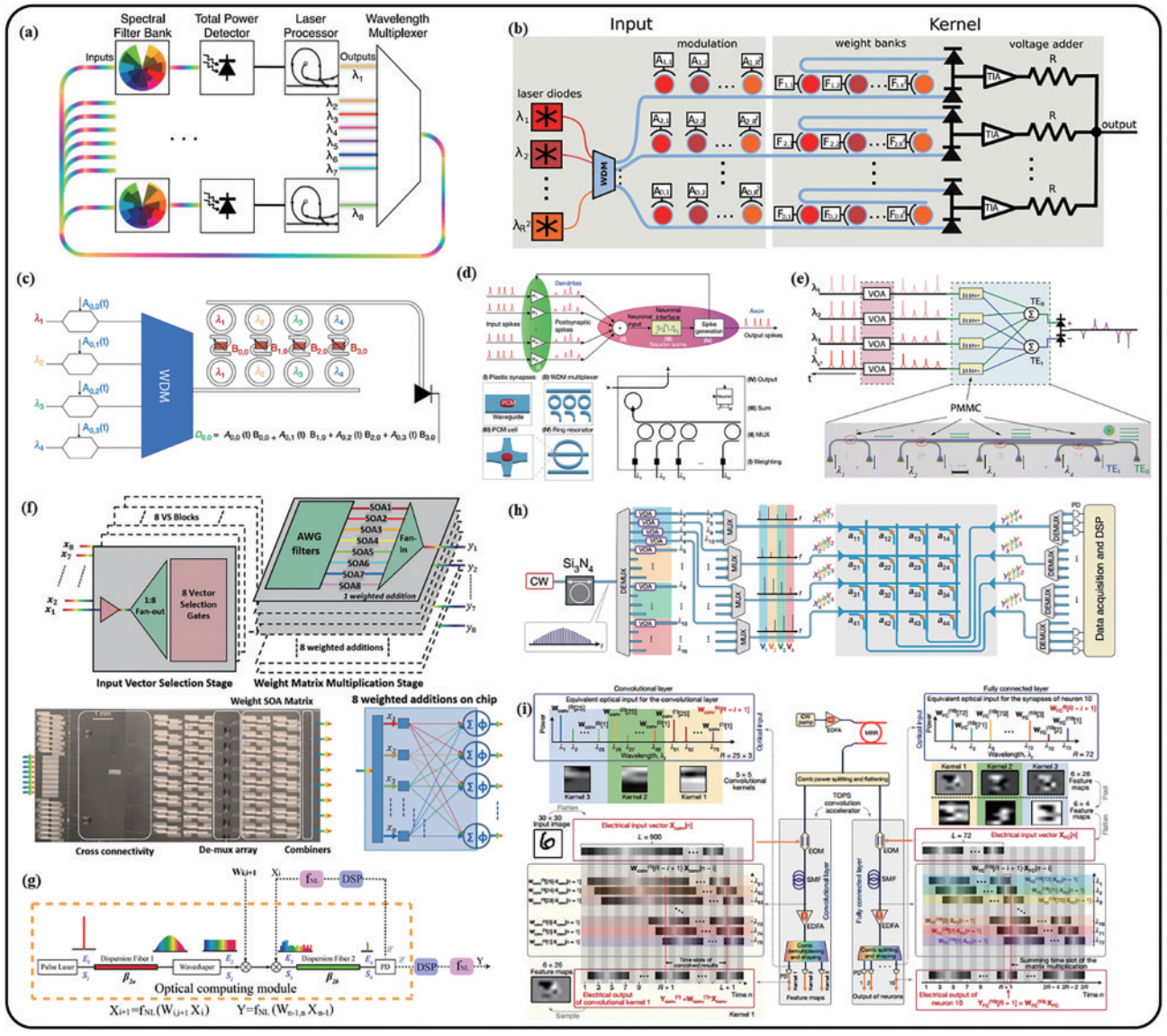

图 7　基于 WDM 技术的光学线性矩阵计算。(a)(b)并联级联的 MRRs[46,89,122]；(c)～(e) PCMs[51,81,92]；(f) SOA[136]；(g)色散光纤[137]；(h)(i)光频梳[93-94]

Fig. 7　Optical linear matrix multiplications based on WDM optics. (a) (b) The cascaded MRRs[46,89,122]; (c)–(e) PCMs[51,81,92]; (f) SOA[136]; (g) dispersion fiber[137]; (h) (i) optical frequency combs[93-94]

Pernice 课题组[93]于 2021 年发表的基于微环光频梳和 PCM 的高度并行全光卷积神经网络。由高 Q 值氮化硅微环产生光频梳，提供架构所需的高度并行的多波长输入，并通过 VOA 实现输入向量的编码；$m \times n$ 规模的片上 MAC 计算由波导上集成的非易失 PCM 实现权值配置。该架构每秒可完成 10^{12} 次 MAC 计算。同年，澳大利亚斯威本科技大学 Moss 课题组[94]发表了另一种基于微环光频梳的光学矢量卷积加速器，如图 7(i)所示。在该架构中，微腔光频梳的输出通过电光马赫-曾德尔调制器(EOM)实现幅度调制，进而在光纤中实现波长依赖的色散时延，从而完成矩阵向量乘法或卷积计算。该方案的算力超过 10 TOPS，即每秒完成 10^{12} 次 MAC 计算，可以实现包含 2.5×10^5 个像素的图像卷积计算。

综上所述，目前已报道的光学线性矩阵计算的工作原理和实现方式包括了基于空间光结构、片上相干原理和 WDM 技术等 3 大类，本综述对这 3 类实现方式进行了比较与分析，如表 1 所示。其中，基于空间光结构的光学线性矩阵计算充分利用其光学三维互连能力，发挥了光在自由空间中的高度并行性，进而通过不同的空间介质实现对光场的幅度和相位调制。但也存在着构成元件加工精度限制、不易集成、部分方案不可编程控制等问题，并在成本和电子 CMOS(如内存存储、带宽等)的协同上面临着巨大挑战。基于片上相干原理和 WDM 技术的架构方案具有可高度集成、可编程控制、硬件可拓展和可重构等特点。基于片上相干原理的光学线性矩阵计算依赖于矩阵分解，其误差会随着 MZI 拓扑级联而累积，且在大规模集成拓展中受

表 1　不同光学线性矩阵计算实现方式的对比

Table 1　Comparison among different implementations of optical linear matrix multiplication

Implementation	Coherent computing	Integration	Weight configuration	Advantage	Limitation
Based on free-space optics	Both	No	One-one	High parallelism	Manufacture precision, peripheral circuit performance
Based on coherent optics	Yes	Yes	SVD, programmable control	Extensibility and reconfigurability	Error accumulation, wafer size
Based on WDM optics	No	Yes	One-one, programmable control	Extensibility and reconfigurability	Wavelength alignment, system control

到晶圆尺寸的限制。另外，基于 WDM 的架构方案可以实现光学物理单元和权重矩阵元素、输入向量元素之间的一一对应，对算法辅助的依赖较低。但是，其在控制和配置上需要光学物理单元的多波长对齐，这使得该架构在大规模拓展时控制难度急剧增加。正是由于不同实现方式的架构均存在各自的优缺点，当前光神经网络的技术路线百花齐放。

3.2　光学神经网络中非线性激活器的研究现状和发展趋势

对于不含任何非线性激活函数的 ANN 来说，即使物理上存在着多层线性变换，由于多个线性矩阵乘的结果仍是一个单一的矩阵，其有效的计算也只等价于单层[139]。为了让 ANN 具备完善的计算和信号处理能力，非线性函数是其中不可缺少的基本算子。图 8 中包含了 ANN 中常用的 6 种非线性激活函数和其数学表达式。为了实现全光神经网络，非线性激活器的光学实现同样必不可少。然而，在目前报道的光学神经网络架构研究中，非线性函数通常是在电上实现的，如 CPU 或 DSP 上，这需要额外引入数模/模数转换器件、光电/电光转换器件、其他外围驱动电路、等等，无疑会增加系统的复杂度、时延、功耗。因此，如何实现低功耗、低非线性产生阈值、高响应速率的片上集成、全光非线性激活器成为了光学神经网络研究中极具挑战的问题之一。

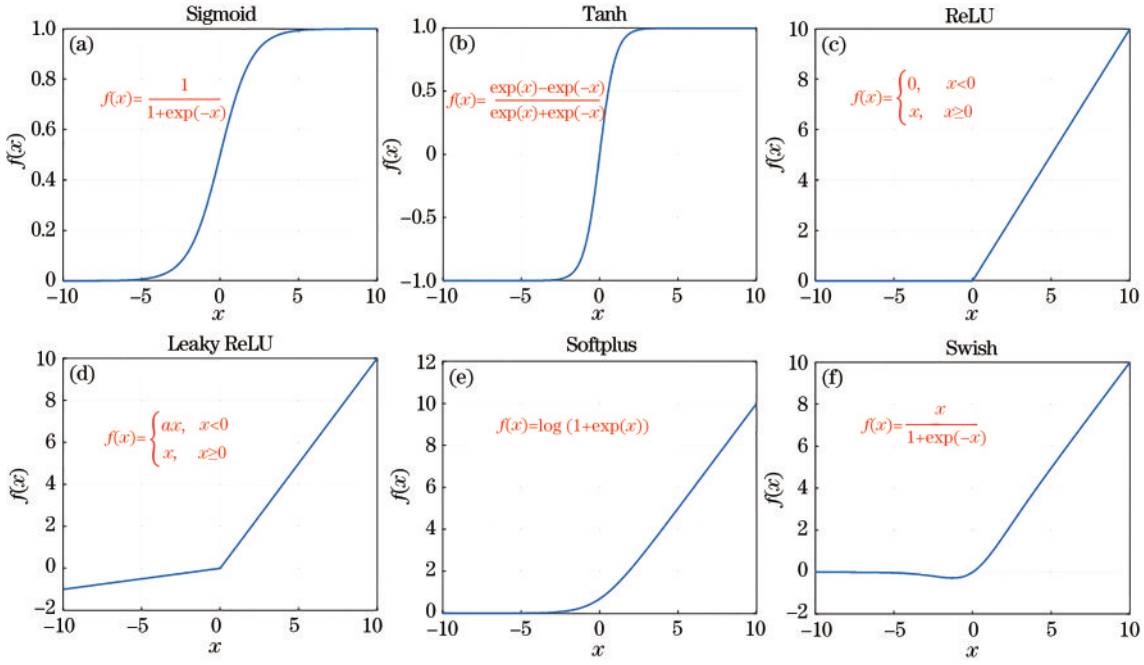

图 8　ANN 中常用的非线性激活函数。(a) Sigmoid；(b) Tanh；(c) ReLU；(d) Leaky ReLU；(e) Softplus；(f) Swish

Fig. 8　Typical expressions of nonlinear activation functions in ANNs. (a) Sigmoid; (b) Tanh; (c) Relu; (d) Leaky Relu; (e) Softpuls; (f) Swish

目前已报道的光学非线性激活函数器主要包括基于光-电-光(O-E-O)转换[140-143]和全光[51,76,123,144-152]两种方式。

3.2.1　基于 O-E-O 转换的光学非线性激活器

2019 年，乔治华盛顿大学 Sorger 课题组[140]发表了基于电吸收调制器(EAM)的非线性激活器，如图 9(a)所示。该工作提取了 5 种不同类型 EAM 的饱和曲线，并比较其产生的不同非线性函数在光学神经网络中的性能，证明基于量子阱的 EAM 在手写数字识别应用上能达到 96% 的准确率，且功耗水平维持在 1.7×10^{-12} J/MAC。同年，该课题组[141]报道了基于氧化铟锡(ITO)的 EAM，如图 9(b)所示。该器件同硅基光子平台相兼容，所提取的非线性激活函数在手写数字识别上能达到 97% 的准确率。另外，普林斯顿大学

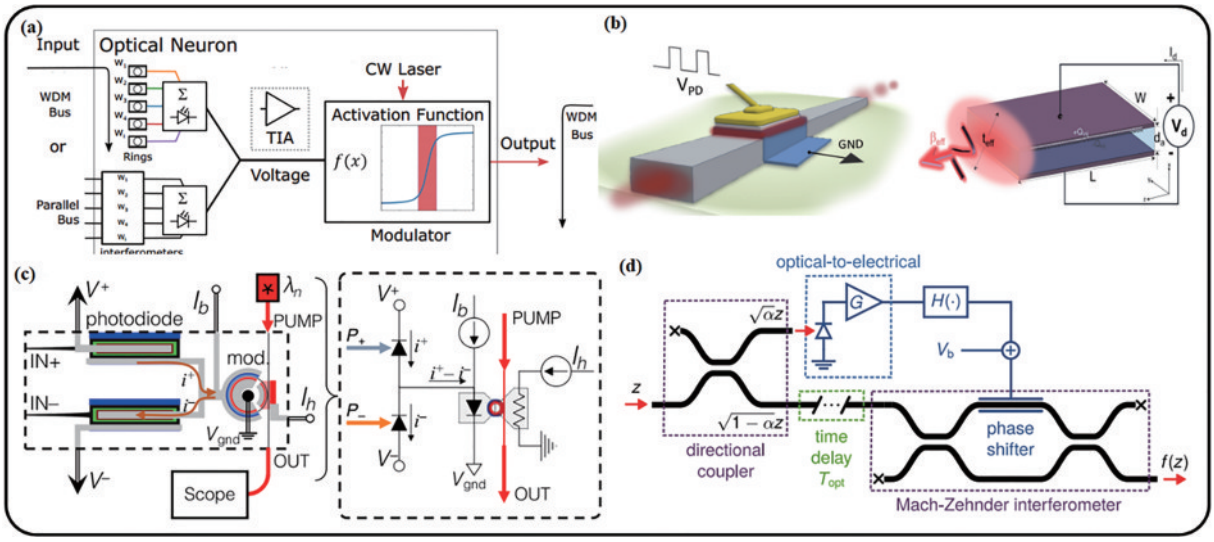

图 9　基于 O-E-O 转换的光学非线性激活器。(a)(b)电光调制器[140-141]；(c)微环调制器[142]；(d)电光辅助 MZI[143]

Fig. 9　Optical nonlinear activators based on O-E-O conversion. (a)(b) Electro-optic modulators[93-94]; (c) MRR modulator[93-94]; (d) feedback-assisted MZI[143]

Prucnal 课题组[142]发表了基于 MRR 调制器实现的非线性激活函数，如图 9(c)所示。该结构所产生的不同非线性响应曲线由泵浦光波长和 MRR 谐振波长之间的波长偏移量决定。2020 年，斯坦福大学和 GenXComm 公司联合发表了基于电光辅助 MZI 的光到光非线性激活器[143]，如图 9(d)所示。该方案由 1∶99 的定向耦合器(DC)和 MZI 构成，依据所建立的目标非线性函数中光强和 MZI 调制电压的查找表，通过 DC 耦合得到的 1% 光能量来确定原光路上的非线性取值所需的 MZI 调制电压值，从而实现特定的非线性函数类型。

上文所述的通过 O-E-O 转换实现的光学非线性激活器，仍存在的光电信号互相转换会不可避免地降低光学神经网络的计算速度，增加系统功耗与延时。理想的方案是在全光情况下实现低功耗、低非线性产生阈值以及高响应速度的非线性激活器，并最终集成在光学线性矩阵计算芯片内。

3.2.2　全光非线性激活器

目前已报道的全光非线性激活器包括基于定制化材料[51,76,123,144-145]、SOA[146-148]以及 MRR[149-152]等 3 大类，如图 10 所示。其中，图 10(a)～(f)是 6 种具有代表性的基于定制化材料的全光非线性激活器。图 10(a)～(c)分别是饱和吸收体[144]、反向饱和吸收体[145]和基于电磁诱导透明(EIT)的腔体[76]及光功率响应曲线，这 3 种结构通常适用于自由空间中的全光非线性激活。图 10(d)为一对金纳米粒子中嵌入单量子点的结构[145]。该结构可与硅光波导集成，同时其在波导中集成的位置不同能产生不同形状的非线性函数。另外，图 10(e)、(f)为两种基于 PCM 的 CMOS 工艺兼容的全光非线性激活器[51,123]。

图 10(g)～(i)是 3 种基于 SOA 的全光非线性激活器[146-148]。荷兰埃因霍温大学 Stabile 课题组[146]于 2019 年报道了基于单个非线性 SOA 的波长变换功能实现的全光非线性函数，如图 10(g)所示。非线性 SOA 可以将求和后的多波长信息转换到新的波长输入上，这一幅度转换过程可以拟合为 Tanh 函数，同时可以通过改变 SOA 的驱动电流来调制非线性函数的斜率，从而实现所需的非线性函数。同年，希腊塞萨洛尼基亚里士多德大学 Mourgias-Alexandris 课题组[148]发表了一种差分偏置的 SOA-MZI 和 SOA 交叉增益调制门的组合结构，该结构可以实现 Sigmoid 函数，如图 10(h)所示。该结构能用于基于 WDM 技术的光学线性矩阵计算架构中，实验演示了其在四波长并行的权重加权求和后实现的全光非线性计算能力。2021 年，塞尔维亚贝尔格莱德大学 Gvozdić 课题组[147]发表了基于注入锁定(injection-locked)的法布里-珀罗激光二极管(FP-LD)结构实现的类 Sigmoid 和 PReLU 非线性函数，如图 10(i)所示。该结构利用 LD 的双稳态特性，并通过调节输入信号和注入锁定边模之间存在的角频率失谐量实现两种不同形状的非线性函数。

另外，图 10(j)～(l)展示了 3 种基于 MRR 结构且 CMOS 工艺兼容的全光非线性激活器[149-152]，提供了光学非线性函数与光学线性矩阵计算的在光子集成回路上的实现可能性，有望进一步实现全光神经网络。其中，图 10(j)是普林斯顿大学 Prucnal 课题组[151-152]于 2020 年提出的可重构全光非线性激活器。该器件基于 MRR 中存在的载流子色散效应，设计了可编程的 MZI 辅助型 MRR 结构，该结构通过调节输入光的耦合比以及输入光波长和 MRR 谐振波长的偏移量，可配置成 4 种不同的非线性函数，从而满足不同的光学神经网络需求。进而，该课题组[149]于 2022 年发表了一种基于硅上石墨烯结构的非线性 MRR，如图 10(k)所

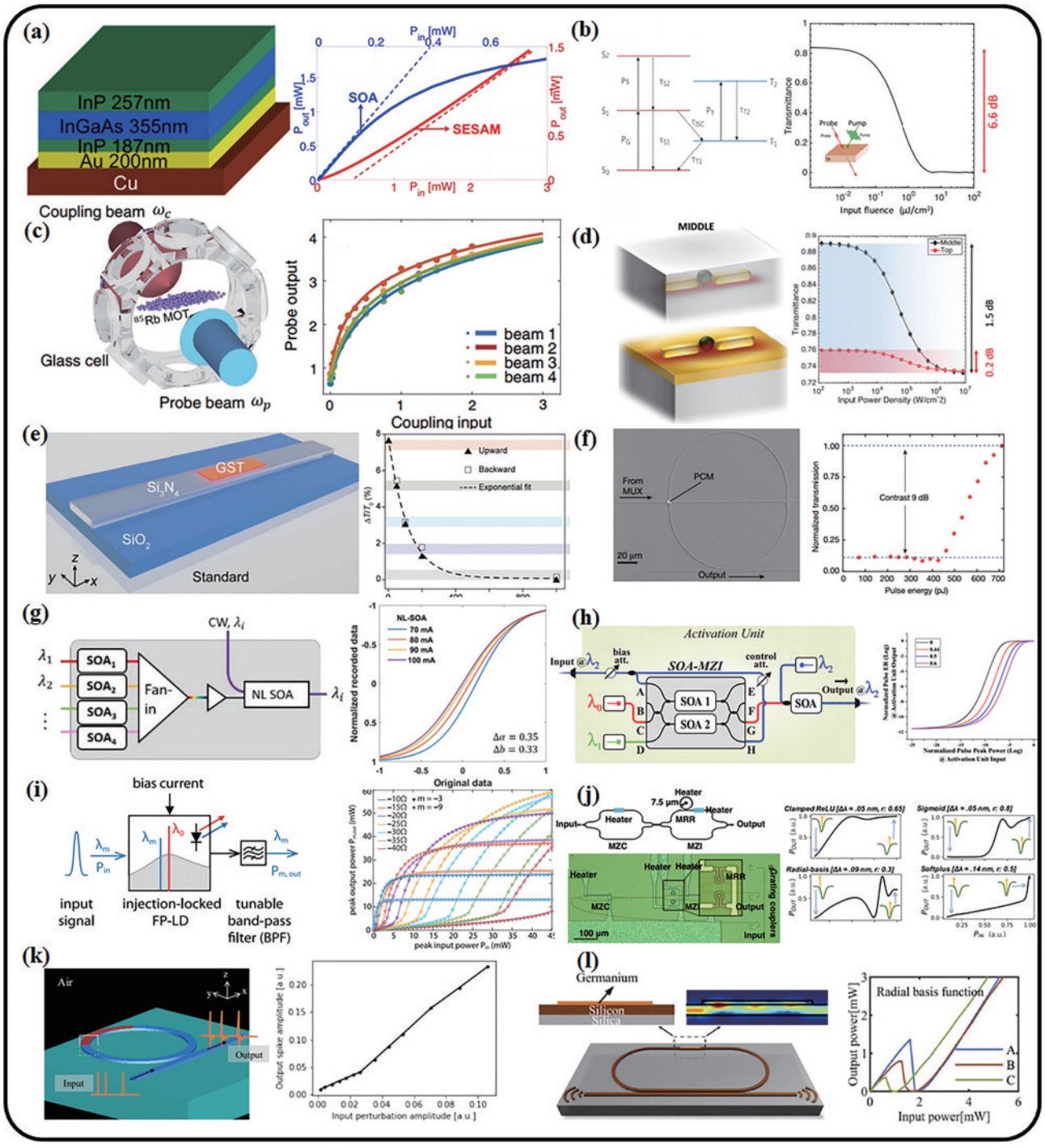

图 10　全光非线性激活器及其响应曲线。(a)～(f)定制化材料[51,76,123,144-145];(g)～(i) SOAs[146-148];(j)～(l) MRRs[149-152]

Fig. 10　All-optical nonlinear activators and the corresponding response curves. (a)–(f) Custom-defined materials[51,76,123,144-145]; (g)–(i) SOAs[146-148]; (j)–(l) MRRs[149-152]

示。得益于石墨烯材料更强的非线性效应,该结构的响应速度可以达到GHz量级。同年,华中科技大学张新亮课题组[150]发表了基于锗硅(Ge/Si)混合材料的跑道型MRR的全光非线性激活器,如图10(l)所示。基于Ge在1550 nm波长附近具有较大的热光系数,该方案实现了响应阈值为0.75 mW的3种不同非线性函数,从而实现了一种CMOS工艺兼容的、低非线性产生阈值的超紧凑全光非线性激活器。

3.3　光学神经网络系统架构与应用的研究现状和发展趋势

进一步地,梳理了光学神经网络系统的完整设计流程,如图11所示。光学神经网络系统的设计与实现,覆盖了从底层的光学元器件、光学大规模阵列架构、光电混合或全光系统架构、软硬件协同乃至目标应用场景的全流程设计与实现。按照从应用层出发,逐渐延展至物理层的设计逻辑,其自上而下可分为目标

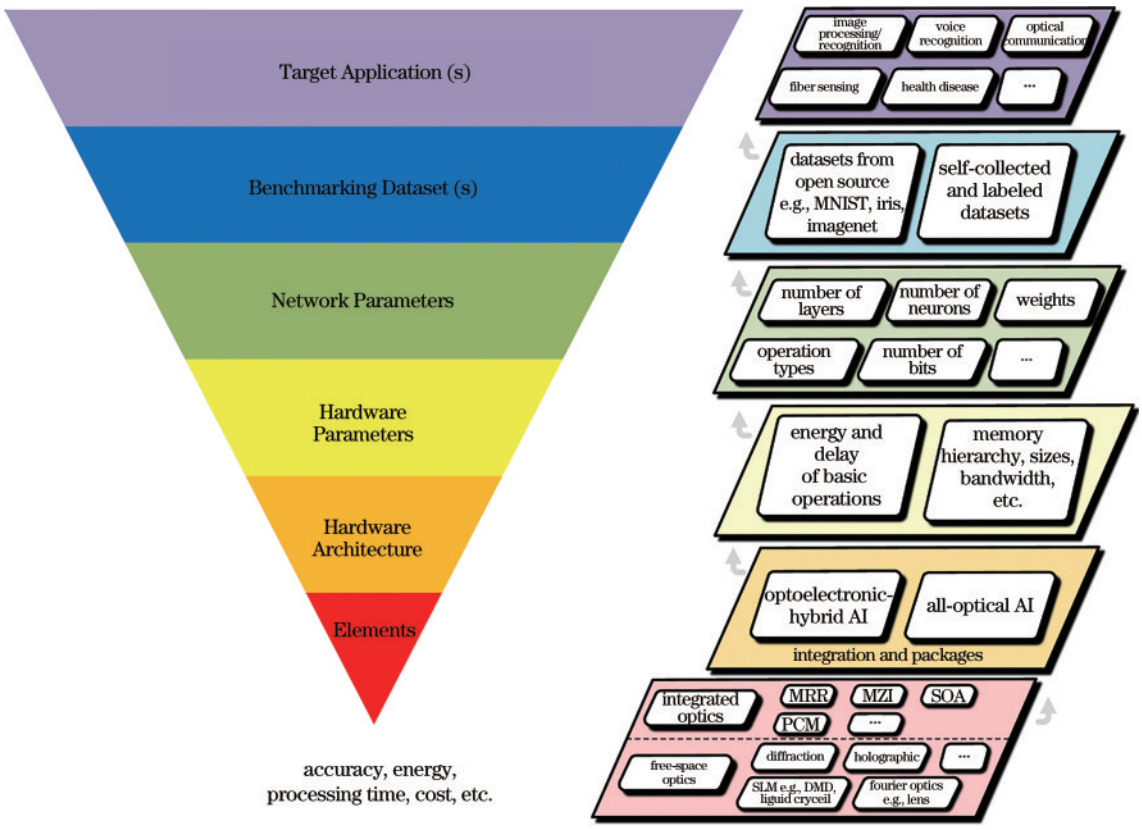

图 11　光学神经网络系统的完整设计流程

Fig. 11　Design process of photonic neural networks

应用确定、数据集标定、网络参数训练与性能评估、硬件参数确定与性能评估、硬件架构设计、基本光学算子实现等 6 个方面。其中，每一方面所需考虑的要素都整理并归纳于图 11 中，可为后续光神经网络系统的设计与实现提供方法论上的指导。其中，对于光学神经网络芯片而言，本综述总结其 3 大重要性能指标为单次光学矩阵计算的延时、单次光学矩阵计算所需功耗以及单位面积上每秒能实现的计算量。这 3 个指标可用于对比并评价不同架构的光学神经网络芯片的计算性能。

众所周知，蓬勃发展的人工智能技术已在机器视觉、自动驾驶、棋盘游戏和临床诊断等各个领域得到了变革性的应用，充分展示了其强大的能力和澎湃的活动。光学神经网络作为一种新型的 AI 计算硬件架构，其应用场景也正从图像领域逐渐拓展到更广泛的领域中去。本综述归纳总结了目前已报道的光学神经网络的应用场景，如图 12 所示。其中，图 12(a)～(e)展示了光学神经网络在图像领域中的一些典型应用，围绕图像的处理和识别，分别为 MNIST 手写数字识别[93-94]、图像边缘检测[92]、鸢尾花(Iris)识别[136]、图像非线性分类[82]、医疗图像恢复[134]。另外，图 12(f)～(i)也展示了光学神经网络在元音识别[73]、轨道角动量(OAM)的复用和解复用[153]、光学逻辑门计算[80]以及光纤的非线性补偿[154]中的应用。这些应用充分展示并验证了：当前不断涌现的光学神经网络架构在人工智能的特定应用上，其处理速度能达到数量级的提升，同时也在计算功耗上带来数量级的降低。

尽管如此，目前已报道的光计算及光神经网络系统大多仍以光电混合框架的形式存在。而在综合评估光电混合的计算架构性能时，其整机系统在处理速度、功耗、计算量等性能参数上，相较于理想的全光计算架构来说，均有所下降。因此，无论学术界还是工业界仍长期致力于研究并制备出全光的计算系统，以充分发挥光传输所具备的高通量、低延迟、低能耗等优势，为新型的高性能计算硬件处理器提供高速、可靠、可行、灵活的方案。而对于光计算以及光神经网络系统来说，全光架构仍面临着巨大挑战，也存在着诸多待解决的问题，例如：如何实现低功耗、低非线性产生阈值、高响应速度的片上集成全光非线性计算单元？如何在模拟的全光计算中实现高精度数值控制？如何将不同功能的光学算子集成到单芯片中？如何在芯片中实现低功耗、灵活可重构的高速光学算子？等等。

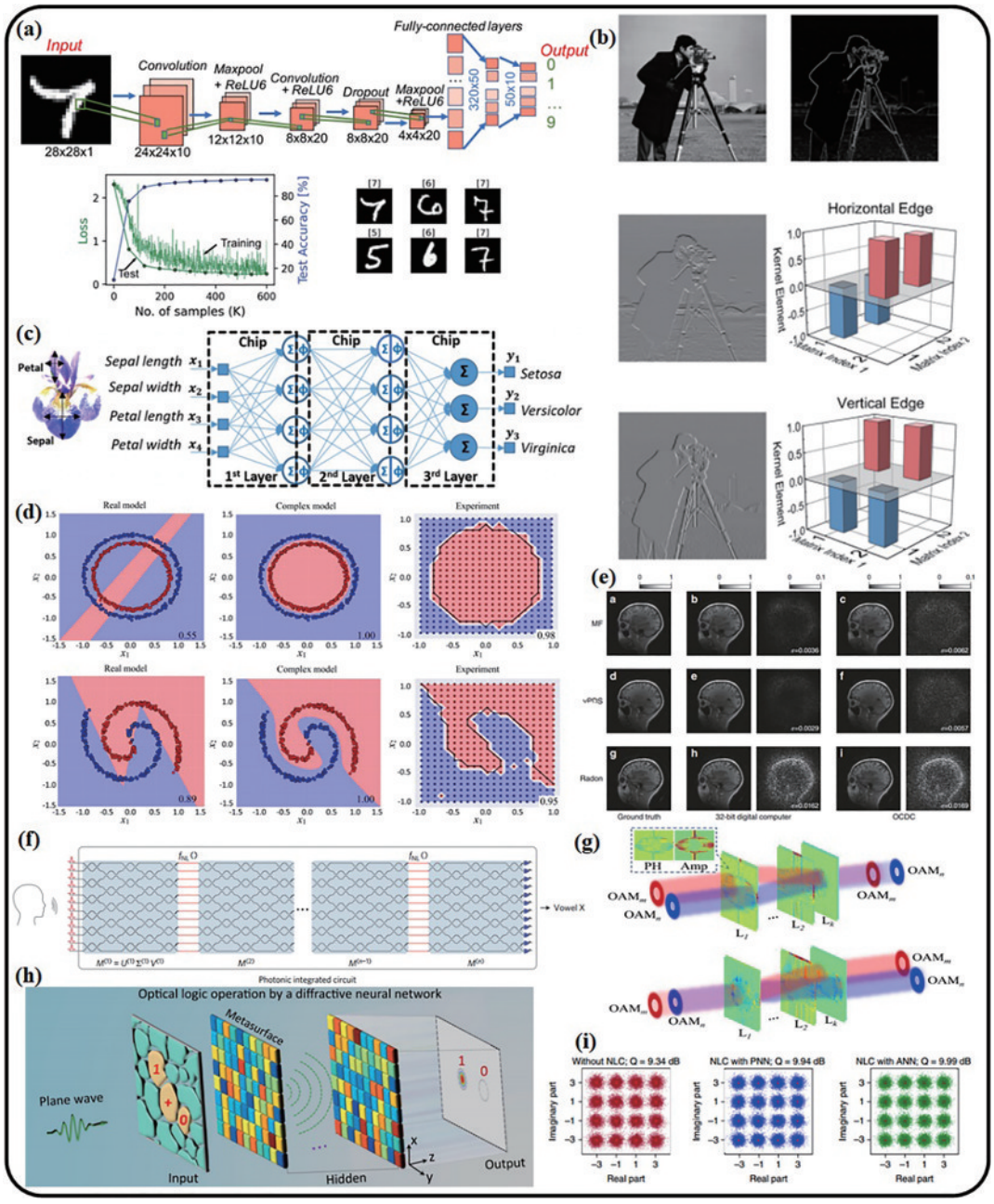

图 12　光学神经网络的典型应用。(a)～(e)图像处理和图像识别[82,92-94,134,136]；(f)元音识别[73]；(g) OAM 的复用与解复用[153]；(h)逻辑门运算[80]；(i)光纤非线性补偿[154]

Fig. 12　Typical applications of photonic neural networks. (a)–(e) Image processing or recognition[82,92-94,134,136]; (f) vowel recognition[73]; (g) OAM multiplexing and demultiplexing[153]; (h) logic operation[80]; (i) fiber nonlinearity compensation[154]

4　结束语

自 2017 年研究人员直接利用光在传输过程中存在的光场变化实现特定计算功能的模拟光计算后，“传输即计算”的概念被提出。光计算及光神经网络成为光电子学、微电子学、数学、算法以及计算机系统等深度融合的交叉学科研究方向，其主要研究内容包括光计算及光神经网络的架构创新、关键光学算子(如光学线性矩阵计算、光学卷积、光学非线性激活等)的实现、片上训练、物理层适配的光计算算法革新、系统应用探索，以及相关软件与生态构建，等等。对于模拟运算的光计算及光神经网络来说，其天然存在着系统噪声，在精度上存在物理器件的限制，如何在低精度和噪声累积的条件下实现高准确率的计算成为挑战。同时，在探索具有高性能的光电混合计算硬件系统或全光计算硬件系统的架构创新上，两者需并驾齐驱。另外，针对系统的实际能效提升，需在同一种应用下用统一维度对光电系统进行公平的能效对比。随着交叉学科中不同领域研究者的共同努力，光计算及光神经网络能实现更为广泛的应用。

参 考 文 献

[1] McCarthy J, Minsky M L, Rochester N, et al. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955[J]. AI Magazine, 2006, 27(4): 12-14.

[2] Leijnen S, van Veen F. The neural network zoo[J]. *Proceedings*, 2020, 47(1): 9.

[3] Ouyang J, Wu E, Wang J, et al. XPU: a programmable FPGA accelerator for diverse workloads[EB/OL]. [2022-03-06]. https://old. hotchips. org/wp-content/uploads/hc_archives/hc29/HC29.21-Monday-Pub/HC29.21.40-Processors-Pub/HC29.21.410-XPU-FPGA-Ouyang-Baidu.pdf.

[4] Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit[EB/OL]. (2017-04-16) [2022-03-06]. https://arxiv. org/abs/1704.04760.

[5] Chen T S, Du Z D, Sun N H, et al. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning[J]. ACM SIGARCH Computer Architecture News, 2014, 42(1): 269-284.

[6] Zhang S J, Du Z D, Zhang L, et al. Cambricon-X: an accelerator for sparse neural networks[C]//2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), October 15-19, 2016, Taipei, China. New York: IEEE Press, 2016.

[7] Merolla P A, Arthur J V, Alvarez-Icaza R, et al. Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface[J]. Science, 2014, 345(6197): 668-673.

[8] Pei J, Deng L, Song S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture[J]. Nature, 2019, 572(7767): 106-111.

[9] Lohn A J, Musser M. How much longer can computing power, drive artificial intelligence progress? [EB/OL]. [2022-10-08]. https://cset. georgetown. edu/wp-content/uploads/AI-and-Compute-How-Much-Longer-Can-Computing-Power-Drive-Artificial-Intelligence-Progress_v2.pdf.

[10] Hernandez D, Brown T B. Measuring the algorithmic efficiency of neural networks[EB/OL]. (2020-05-08) [2022-03-06]. https://arxiv.org/abs/2005.04305.

[11] Waldrop M M. More than moore[J]. Nature, 2016, 530(7589): 144-148.

[12] Ambs P. Optical computing: a 60-year adventure[J]. Advances in Optical Technologies, 2010, 2010: 372652.

[13] Kneale W. Iv. Boole and the revival of logic[J]. Mind, 1948, LVII(226): 149-175.

[14] Jain K, Pratt G W. Optical transistor[J]. Applied Physics Letters, 1976, 28(12): 719-721.

[15] Athale R A, Lee S H. Development of an optical parallel logic device and a half-adder circuit for digital optical processing[J]. Optical Engineering, 1979, 18(5): 513-517.

[16] Jenkins B K, Sawchuk A A, Strand T C, et al. Sequential optical logic implementation[J]. Applied Optics, 1984, 23(19): 3455-3464.

[17] Tanida J, Ichioka Y. Optical-logic-array processor using shadowgrams III Parallel neighborhood operations and an architecture of an optical digital-computing system[J]. Journal of the Optical Society of America A, 1985, 2(8): 1245-1253.

[18] Tanida J, Ichioka Y. OPALS: optical parallel array logic system[J]. Applied Optics, 1986, 25(10): 1565-1570.

[19] Main T, Feuerstein R J, Jordan H F, et al. Implementation of a general-purpose stored-program digital optical computer[J]. Applied Optics, 1994, 33(8): 1619-1628.

[20] Miller D, Smith S, Seaton C. Optical bistability in semiconductors[J]. IEEE Journal of Quantum Electronics, 1981, 17(3): 312-317.

[21] Miller D A B. Are optical transistors the logical next step? [J]. Nature Photonics, 2010, 4(1): 3-5.

[22] Zhuang L M, Roeloffzen C G H, Hoekman M, et al. Programmable photonic signal processor chip for radiofrequency applications[J]. Optica, 2015, 2(10): 854-859.

[23] Pérez D, Gasulla I, Capmany J, et al. Reconfigurable lattice mesh designs for programmable photonic processors [J]. Optics Express, 2016, 24(11): 12093-12106.

[24] Liu W L, Li M, Guzzon R S, et al. A fully reconfigurable photonic integrated signal processor[J]. Nature Photonics, 2016, 10(3): 190-195.

[25] Pérez D, Gasulla I, Crudgington L, et al. Multipurpose silicon photonics signal processor core[J]. Nature Communications, 2017, 8: 636.

[26] Perez D, Gasulla I, Fraile F J, et al. Silicon photonics rectangular universal interferometer[J]. Laser & Photonics Reviews, 2017, 11(6): 1700219.

[27] Bogaerts W, Pérez D, Capmany J, et al. Programmable photonic circuits[J]. Nature, 2020, 586(7828): 207-216.

[28] Zhou H L, Zhao Y H, Wang X, et al. Self-configuring and reconfigurable silicon photonic signal processor[J]. ACS Photonics, 2020, 7(3): 792-799.

[29] Pérez-López D, Gutiérrez A, Capmany J. Silicon nitride programmable photonic processor with folded heaters[J]. Optics Express, 2021, 29(6): 9043-9059.

[30] Wang Z, Marandi A, Wen K, et al. Coherent Ising machine based on degenerate optical parametric oscillators [J]. Physical Review A, 2013, 88(6): 063853.

[31] Marandi A, Wang Z, Takata K, et al. Network of time-multiplexed optical parametric oscillators as a coherent Ising machine[J]. Nature Photonics, 2014, 8(12): 937-942.

[32] McMahon P L, Marandi A, Haribara Y, et al. A fully programmable 100-spin coherent Ising machine with all-to-all connections[J]. Science, 2016, 354(6312): 614-617.

[33] Inagaki T, Inaba K, Hamerly R, et al. Large-scale Ising spin network based on degenerate optical parametric oscillators[J]. Nature Photonics, 2016, 10(6): 415-419.

[34] Takesue H, Inagaki T. 10 GHz clock time-multiplexed degenerate optical parametric oscillators for a photonic Ising spin network[J]. Optics Letters, 2016, 41(18): 4273-4276.

[35] Yamamoto Y, Aihara K, Leleu T, et al. Coherent Ising machines: optical neural networks operating at the

quantum limit[J]. Npj Quantum Information, 2017, 3: 49.
[36] Takesue H, Inagaki T, Inaba K, et al. Large-scale coherent Ising machine[J]. Journal of the Physical Society of Japan, 2019, 88(6): 061014.
[37] Hamerly R, Inagaki T, McMahon P L, et al. Experimental investigation of performance differences between coherent Ising machines and a quantum annealer [J]. Science Advances, 2019, 5(5): eaau0823.
[38] Cen Q Z, Hao T F, Ding H, et al. Microwave photonic Ising machine[EB/OL]. (2020-09-19)[2022-03-05]. https://arxiv.org/abs/2011.00064.
[39] Böhm F, Verschaffelt G, Van der Sande G. A poor man's coherent Ising machine based on opto-electronic feedback systems for solving optimization problems[J]. Nature Communications, 2019, 10(1): 3538.
[40] Babaeian M, Nguyen D T, Demir V, et al. A single shot coherent Ising machine based on a network of injection-locked multicore fiber lasers[J]. Nature Communications, 2019, 10: 3516.
[41] Pierangeli D, Marcucci G, Conti C. Large-scale photonic Ising machine by spatial light modulation[J]. Physical Review Letters, 2019, 122(21): 213902.
[42] Pierangeli D, Marcucci G, Conti C. Adiabatic evolution on a spatial-photonic Ising machine[J]. Optica, 2020, 7 (11): 1535-1543.
[43] Prabhu M, Roques-Carmes C, Shen Y C, et al. Accelerating recurrent Ising machines in photonic integrated circuits[J]. Optica, 2020, 7(5): 551-558.
[44] Okawachi Y, Yu M J, Jang J K, et al. Demonstration of chip-based coupled degenerate optical parametric oscillators for realizing a nanophotonic spin-glass[J]. Nature Communications, 2020, 11: 4119.
[45] Tait A N, Nahmias M A, Tian Y, et al. Photonic neuromorphic signal processing and computing[M]// Naruse M. Nanophotonic information physics. Nano-optics and nanophotonics. Heidelberg: Springer, 2013: 183-222.
[46] 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): 3427-3439.
[47] Shastri B J, Nahmias M A, Tait A N, et al. Spike processing with a graphene excitable laser[J]. Scientific Reports, 2016, 6: 19126.
[48] Nahmias M A, Peng H T, de Lima T F, et al. A TeraMAC neuromorphic photonic processor[C]//2018 IEEE Photonics Conference, September 30-October 4, 2018, Reston, VA, USA. New York: IEEE Press, 2018.
[49] Chakraborty I, Saha G, Sengupta A, et al. Toward fast neural computing using all-photonic phase change spiking neurons[J]. Scientific Reports, 2018, 8: 12980.
[50] Shainline J M, Buckley S M, McCaughan A N, et al. Circuit designs for superconducting optoelectronic loop neurons[J]. Journal of Applied Physics, 2018, 124(15): 152130.
[51] 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.
[52] Mehrabian A, Sorger V J, El-Ghazawi T. A design methodology for post-Moore's law accelerators: the case of a photonic neuromorphic processor[C]//2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors, July 6-8, 2020, Manchester, UK. New York: IEEE Press, 2020: 113-116.
[53] Miscuglio M, Meng J W, Yesiliurt O, et al. Artificial synapse with mnemonic functionality using GSST-based photonic integrated memory[C]//2020 International Applied Computational Electromagnetics Society Symposium (ACES), July 27-31, 2020, Monterey, CA, USA. New York: IEEE Press, 2020.
[54] Skontranis M, Sarantoglou G, Deligiannidis S, et al. Time-multiplexed spiking convolutional neural network based on VCSELs for unsupervised image classification [J]. Applied Sciences, 2021, 11(4): 1383.
[55] Appeltant L, Soriano M C, Van der Sande G, et al. Information processing using a single dynamical node as complex system[J]. Nature Communications, 2011, 2: 468.
[56] Vandoorne K. Photonic reservoir computing with a network of coupled semiconductor optical amplifiers[D]. Ghent: Ghent University, 2011.
[57] Vandoorne K, Dambre J, Verstraeten D, et al. Parallel reservoir computing using optical amplifiers[J]. IEEE Transactions on Neural Networks, 2011, 22(9): 1469-1481.
[58] Paquot Y, Duport F, Smerieri A, et al. Optoelectronic reservoir computing[J]. Scientific Reports, 2012, 2: 287.
[59] Fiers M A A, van Vaerenbergh T, Wyffels F, et al. Nanophotonic reservoir computing with photonic crystal cavities to generate periodic patterns[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(2): 344-355.
[60] Vandoorne K, Mechet P, van Vaerenbergh T, et al. Experimental demonstration of reservoir computing on a silicon photonics chip[J]. Nature Communications, 2014, 5: 3541.
[61] Vinckier Q, Duport F, Smerieri A, et al. High-performance photonic reservoir computer based on a coherently driven passive cavity[J]. Optica, 2015, 2(5): 438-446.
[62] Bueno J, Brunner D, Soriano M C, et al. Conditions for reservoir computing performance using semiconductor lasers with delayed optical feedback[J]. Optics Express, 2017, 25(3): 2401-2412.
[63] Katumba A, Freiberger M, Bienstman P, et al. A multiple-input strategy to efficient integrated photonic reservoir computing[J]. Cognitive Computation, 2017, 9 (3): 307-314.
[64] Freiberger M, Katumba A, Bienstman P, et al. Training passive photonic reservoirs with integrated optical readout [J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(7): 1943-1953.
[65] Coarer F D L, Sciamanna M, Katumba A, et al. All-optical reservoir computing on a photonic chip using

silicon-based ring resonators[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2018, 24(6): 7600108.

[66] Katumba A, Heyvaert J, Schneider B, et al. Low-loss photonic reservoir computing with multimode photonic integrated circuits[J]. Scientific Reports, 2018, 8: 2653.

[67] Katumba A, Freiberger M, Laporte F, et al. Neuromorphic computing based on silicon photonics and reservoir computing[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2018, 24(6): 8300310.

[68] Brunner D, Soriano M C, van der Sande G. Photonic reservoir computing[M]. Berlin: De Gruyter, 2019, 8: 19.

[69] Laporte F, Dambre J, Bienstman P. Highly parallel simulation and optimization of photonic circuits in time and frequency domain based on the deep-learning framework PyTorch[J]. Scientific Reports, 2019, 9(1): 5918.

[70] Tanaka G, Yamane T, Héroux J B, et al. Recent advances in physical reservoir computing: a review[J]. Neural Networks, 2019, 115: 100-123.

[71] Mesaritakis C, Syvridis D. Reservoir computing based on transverse modes in a single optical waveguide[J]. Optics Letters, 2019, 44(5): 1218-1221.

[72] Laporte F. Novel architectures for brain-inspired photonic computers[D]. Ghent: Ghent University, 2020.

[73] Shen Y C, Harris N C, Skirlo S, et al. Deep learning with coherent nanophotonic circuits[J]. Nature Photonics, 2017, 11(7): 441-446.

[74] 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.

[75] Hamerly R, Bernstein L, Sludds A, et al. Large-scale optical neural networks based on photoelectric multiplication[J]. Physical Review X, 2019, 9(2): 021032.

[76] Zuo Y, Li B H, Zhao Y J, et al. All-optical neural network with nonlinear activation functions[J]. Optica, 2019, 6(9): 1132-1137.

[77] Shi B, Calabretta N, Stabile R. Image classification with a 3-layer SOA-based photonic integrated neural network [C]//2019 24th OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC), July 7-11, 2019, Fukuoka, Japan. New York: IEEE Press, 2019.

[78] Abel S, Horst F, Stark P, et al. Silicon photonics integration technologies for future computing systems [C]//2019 24th OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC), July 7-11, 2019, Fukuoka, Japan. New York: IEEE Press, 2019.

[79] Zhang T, Wang J, Dan Y H, et al. Efficient training and design of photonic neural network through neuroevolution [J]. Optics Express, 2019, 27(26): 37150-37163.

[80] 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.

[81] Miscuglio M, Sorger V J. Photonic tensor cores for machine learning[J]. Applied Physics Reviews, 2020, 7 (3): 031404.

[82] Zhang H, Gu M, Jiang X D, et al. An optical neural chip for implementing complex-valued neural network[J]. Nature Communications, 2021, 12: 457.

[83] Bagherian H, Skirlo S, Shen Y C, et al. On-chip optical convolutional neural networks[EB/OL]. (2018-08-09) [2022-03-02]. https://arxiv.org/abs/1808.03303.

[84] Mehrabian A, Al-Kabani Y, Sorger V J, et al. PCNNA: a photonic convolutional neural network accelerator[C]// 2018 31st IEEE International System-on-Chip Conference, September 4-7, 2018, Arlington, VA, USA. New York: IEEE Press, 2018: 169-173.

[85] 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.

[86] Liu W C, Liu W Y, Ye Y C, et al. HolyLight: a nanophotonic accelerator for deep learning in data centers [C]//2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), March 25-29, 2019, Florence, Italy. New York: IEEE Press, 2019: 1483-1488.

[87] Xu S F, Wang J, Wang R, et al. High-accuracy optical convolution unit architecture for convolutional neural networks by cascaded acousto-optical modulator arrays [J]. Optics Express, 2019, 27(14): 19778-19787.

[88] Wagner K H, McComb S. Optical rectifying linear units for back-propagation learning in a deep holographic convolutional neural network[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 7701318.

[89] Bangari V, Marquez B A, Miller H, et al. Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs) [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 7701213.

[90] Mehrabian A, Miscuglio M, Alkabani Y, et al. A winograd-based integrated photonics accelerator for convolutional neural networks[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 6100312.

[91] Ahmed M, Al-Hadeethi Y, Bakry A, et al. Integrated photonic FFT for photonic tensor operations towards efficient and high-speed neural networks[J]. Nanophotonics, 2020, 9(13): 4097-4108.

[92] Wu C M, Yu H S, Lee S, et al. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network[J]. Nature Communications, 2021, 12(1): 96.

[93] Feldmann J, Youngblood N, Karpov M, et al. Parallel convolutional processing using an integrated photonic tensor core[J]. Nature, 2021, 589(7840): 52-58.

[94] 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.

[95] 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.
[96] Woeginger G J. Exact algorithms for NP-hard problems: a survey[M]//Jünger M, Reinelt G, Rinaldi G. Combinatorial optimization: eureka, you shrink! Lecture notes in computer science. Heidelberg: Springer, 2003, 2570: 185-207.
[97] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. Bulletin of Mathematical Biology, 1990, 52(1/2): 99-115.
[98] Hebb D O. The organization of behavior: a neuropsychological theory[M]. London: Psychology Press, 2005.
[99] Rosenblatt F. The perceptron, a perceiving and recognizing automaton project para, report: cornell aeronautical laboratory, cornell aeronautical laboratory [EB/OL]. [2022-03-06]. https://books.google.pl/books.
[100] Widrow B, Hoff M E. Adaptive switching circuits[R]. Stanford: Stanford Univ Ca Stanford Electronics Labs, 1960.
[101] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the National Academy of Sciences of the United States of America, 1982, 79(8): 2554-2558.
[102] Kohonen T. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics, 1982, 43 (1): 59-69.
[103] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.
[104] LeCun Y, Boser B, Denker J, et al. Handwritten digit recognition with a back-propagation network[C]// Advances in Neural Information Processing Systems 2, NIPS Conference, November 27-30, 1989, Denver, Colorado, USA. San Francisco: Morgan Kaufmann, 1990.
[105] 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.
[106] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[107] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[108] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]//NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1, December 3-6, 2012, Lake Tahoe, Nevada, USA. New York: ACM Press, 2012: 1097-1105.
[109] Taigman Y, Yang M, Ranzato M, et al. DeepFace: closing the gap to human-level performance in face verification[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 1701-1708.
[110] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2022-03-04]. https://arxiv.org/abs/1409.1556.
[111] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA. New York: IEEE Press, 2015.
[112] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.
[113] Lugt A V. Signal detection by complex spatial filtering [J]. IEEE Transactions on Information Theory, 1964, 10 (2): 139-145.
[114] 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.
[115] Clements W R, Humphreys P C, Metcalf B J, et al. Optimal design for universal multiport interferometers[J]. Optica, 2016, 3(12): 1460-1465.
[116] Psaltis D, Brady D, Gu X G, et al. Holography in artificial neural networks[M]//Yeh P, Gu C. Landmark papers on photorefractive nonlinear optics. Singapore: World Scientific, 1995: 541-546.
[117] Goodman J W, Leonberger F J, Kung S Y, et al. Optical interconnections for VLSI systems[J]. Proceedings of the IEEE, 1984, 72(7): 850-866.
[118] Farhat N H, Psaltis D, Prata A, et al. Optical implementation of the Hopfield model[J]. Applied Optics, 1985, 24(10): 1469-1475.
[119] Yu F T, Lu T, Yang X, et al. Optical neural network with pocket-sized liquid-crystal televisions[J]. Optics Letters, 1990, 15(15): 863-865.
[120] Jang J S, Shin S G, Yuk S W, et al. Dynamic optical interconnections using holographic lenslet arrays for adaptive neural networks[J]. Optical Engineering, 1993, 32(1): 80-87.
[121] Saxena I F, Fiesler E. Adaptive multilayer optical neural network with optical thresholding[J]. Optical Engineering, 1995, 34(8): 2435-2440.
[122] 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.
[123] Cheng Z G, Ríos C, Pernice W H P, et al. On-chip photonic synapse[J]. Science Advances, 2017, 3(9): e1700160.
[124] Sarle W S. Neural networks and statistical models[EB/OL]. [2022-03-06]. https://sss1. bnu. edu. cn/~pguo/cs229/public_html/ann/pdfs/neural1.pdf.
[125] Sermanet P, Eigen D, Zhang X, et al. OverFeat: integrated recognition, localization and detection using convolutional networks[EB/OL]. (2013-12-21)[2022-03-06]. https://arxiv.org/abs/1312.6229.
[126] Li X Q, Zhang G Y, Huang H H, et al. Performance analysis of GPU-based convolutional neural networks [C]//2016 45th International Conference on Parallel Processing (ICPP), August 16-19, 2016, Philadelphia, PA, USA. New York: IEEE Press, 2016: 67-76.
[127] Graham A D. Obtaining high precision results from low

precision hardware[EB/OL]. (2005-12) [2022-06-01]. https://trace. tennessee. edu/cgi/viewcontent. cgi? article=3274&context=utk_gradthes.

[128] Wang T Y, Ma S Y, Wright L G, et al. An optical neural network using less than 1 photon per multiplication [J]. Nature Communications, 2022, 13(1): 123.

[129] Miscuglio M, Hu Z B, Li S R, et al. Massively-parallel amplitude-only Fourier neural network[J]. Optica, 2020, 7(12): 1812-1819.

[130] 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.

[131] Huggins E. Introduction to Fourier optics[J]. The Physics Teacher, 2007, 45(6): 364-368.

[132] Raeker B O, Grbic A. Compound metaoptics for amplitude and phase control of wave fronts[J]. Physical Review Letters, 2019, 122(11): 113901.

[133] Fang M Y S, Manipatruni S, Wierzynski C, et al. Design of optical neural networks with component imprecisions[J]. Optics Express, 2019, 27(10): 14009-14029.

[134] Xu S F, Wang J, Shu H W, et al. Optical coherent dot-product chip for sophisticated deep learning regression[J]. Light: Science & Applications, 2021, 10: 221.

[135] Lawson C L, Hanson R J. Solving least squares problems [M]. Philadelphia: Society for Industrial and Applied Mathematics, 1995.

[136] Shi B, Calabretta N, Stabile R. Deep neural network through an InP SOA-based photonic integrated cross-connect[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 7701111.

[137] Zang Y B, Chen M H, Yang S G, et al. Electro-optical neural networks based on time-stretch method[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 7701410.

[138] Mahjoubfar A, Churkin D V, Barland S, et al. Time stretch and its applications[J]. Nature Photonics, 2017, 11(6): 341-351.

[139] Leshno M, Lin V Y, Pinkus A, et al. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function[J]. Neural Networks, 1993, 6(6): 861-867.

[140] George J K, Mehrabian A, Amin R, et al. Neuromorphic photonics with electro-absorption modulators[J]. Optics Express, 2019, 27(4): 5181-5191.

[141] Amin R, George J K, Sun S, et al. ITO-based electro-absorption modulator for photonic neural activation function[J]. APL Materials, 2019, 7(8): 081112.

[142] Tait A N, de Lima T F, Nahmias M A, et al. Silicon photonic modulator neuron[J]. Physical Review Applied, 2019, 11(6): 064043.

[143] Fard M M P, Williamson I A D, Edwards M, et al. Experimental realization of arbitrary activation functions for optical neural networks[J]. Optics Express, 2020, 28(8): 12138-12148.

[144] Dejonckheere A, Duport F, Smerieri A, et al. All-optical reservoir computer based on saturation of absorption [J]. Optics Express, 2014, 22(9): 10868-10881.

[145] Miscuglio M, Mehrabian A, Hu Z B, et al. All-optical nonlinear activation function for photonic neural networks [J]. Optical Materials Express, 2018, 8(12): 3851-3863.

[146] Shi B, Calabretta N, Stabile R. First demonstration of a two-layer all-optical neural network by using photonic integrated chips and SOAs[C]//45th European Conference on Optical Communication (ECOC 2019), September 22-26, 2019, Dublin, Ireland. New York: IET, 2019.

[147] Crnjanski J, Krstić M, Totović A, et al. Adaptive sigmoid-like and PReLU activation functions for all-optical perceptron[J]. Optics Letters, 2021, 46(9): 2003-2006.

[148] Mourgias-Alexandris G, Tsakyridis A, Passalis N, et al. An all-optical neuron with sigmoid activation function[J]. Optics Express, 2019, 27(7): 9620-9630.

[149] Jha A, Huang C R, Peng H T, et al. Photonic spiking neural networks and graphene-on-silicon spiking neurons [J]. Journal of Lightwave Technology, 2022, 40(9): 2901-2914.

[150] Wu B, Li H K, Tong W Y, et al. Low-threshold all-optical nonlinear activation function based on a Ge/Si hybrid structure in a microring resonator[J]. Optical Materials Express, 2022, 12(3): 970-980.

[151] Jha A, Huang C R, Prucnal P R. Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics[J]. Optics Letters, 2020, 45(17): 4819-4822.

[152] Huang C R, Jha A, de Lima T F, et al. On-chip programmable nonlinear optical signal processor and its applications[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2021, 27(2): 6100211.

[153] Wang P P, Xiong W J, Huang Z B, et al. Diffractive deep neural network for optical orbital angular momentum multiplexing and demultiplexing[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2022, 28(4): 7500111.

[154] 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.

[155] 杨凌雁, 张林. 光蓄水池神经网络研究进展[J]. 中国激光, 2021, 48(19): 1906001.
Yang L Y, Zhang L. Recent progress in photonic reservoir neural network[J]. Chinese Journal of Lasers, 2021, 48(19): 1906001.

[156] 刘家跃, 张建国, 李创业, 等. 基于储备池计算的激光混沌同步保密通信研究[J]. 中国激光, 2022, 49(18): 1806001.
Liu J Y, Zhang J G, Li C Y, et al. Secure communication via laser chaos synchronization based on reservoir computing[J]. Chinese Journal of Lasers, 2022, 49(18): 1806001.

[157] 刘雅名, 郭宏翔, 陈彦虎, 等. 基于光子计算的随机奇异值分解[J]. 光学学报, 2022, 42(19): 1920002.
Liu Y M, Guo H X, Chen Y H, et al. Random singular value decomposition based on optical computation[J]. Acta Optica Sinica, 2022, 42(19): 1920002.