“2024 年度中国科学十大进展”解读

大规模智能光子计算推理的突破性进展

潘婧 [1†]，李志浩 [1†]，仇旻 [1,2*]

1. 西湖大学工学院电子信息工程系，全省 3D 微纳加工和表征研究重点实验室，杭州 310030
2. 西湖大学光电研究院，杭州 311421

† 同等贡献

* 联系人, E-mail: qiumin@westlake.edu.cn

2025-05-20 收稿, 2025-09-23 修回, 2025-09-23 接受, 2025-09-24 网络版发表

摘要 光学神经网络有着“结构即功能，传播即计算”的独特物理实现方式，在高性能计算领域引发了广泛关注。随着人工智能时代全面到来，算力需求呈现爆发式增长。光子计算凭借高速计算、高并行性以及超低功耗等优势，给全世界范围内的研究人员提供了一种颠覆性的计算架构和全新的技术路径。然而，光学神经网络的训练步骤普遍依赖 GPU 离线训练-参数在线部署过程。如何有效实现在线训练，实时调整网络参数，成为光计算领域急需解决的关键问题。清华大学方璐、戴琼海团队借助全前向智能光计算训练架构，打造出了国际首款大规模通用智能光计算芯片“太极”，实现了大规模神经网络的高效推理与训练。这项工作被评选为“2024 年度中国科学十大进展”，为通用人工智能的高效能计算开拓了一条新路径。本文回顾了光学神经网络的发展历程，及其在架构设计、网络训练和应用领域的最新研究进展，着重分析了“太极”芯片在光计算领域的重要意义，及其在产业化应用中的发展前景。最后，对光计算的未来发展趋势进行了展望。

关键词 光计算，光学神经网络，光计算架构，大规模智能光子计算，光电智能计算，人工智能

人工智能（artificial intelligence, AI）技术的发展最早能追溯至1956年的达特茅斯会议，约翰·麦卡锡（John McCarthy）、克劳德·香农（Claude Shannon）、艾伦·纽厄尔（Allen Newell）等人首次提出“人工智能”这一术语，标志着人工智能的诞生。在接下来的70多年里，人工智能技术经历了多次起伏，直到半导体工艺进步推动算力快速增长[1]、互联网普及引发数据规模爆发、误差反向传播算法突破单层感知机局限，人工智能才彻底摆脱了《莱特希尔报告》对其“错误且徒劳”的负面评价，开启了从实验室到现实应用的技术转化和范式跃迁。在2025年的当下，以深度学习[2]、大模型[3]为代表的先进算法在海量数据资源的驱动下迅猛发展，对计算资源的需求呈爆发式上涨。与此同时，半导体与集成电子工艺快速发展，处理器架构不断刷新性能记录。从通用中央处理器（central processing unit, CPU）到大规模并行计算图形处理器（graphics processing unit, GPU），从可编程门阵列（field programmable gate array, FPGA）动态重构芯片到专用集成电路（application specific integrated circuit, ASIC）定制化加速器，硬件的升级使得千亿参数级模型的训练成为可能。然而，受限于冯·诺依曼“存算分离”的架构体系以及半导体器件的物理极限[4]，传统电子计算芯片的算力增长速度逐渐难以匹配爆发增长的算力需求，新兴智能计算范式的开发迫在眉睫。

光子计算[5-7]是近些年发展起来的新兴计算范式，具有“结构即功能，传播即计算”的独特物理实现方式。光作为光子计算网络的数据载体，借助干涉[8]、衍射[9]等物理过程，在进行矩阵运算、信息传递等计算进程中几乎不消耗能量。并且，光具有波长、相位、幅度、偏振、角动量等调控维度[10]，运用波分复用、频分复用、偏振复用等编码技术能迅速提升网络的容量以及计算带宽。与电子（费米子）不同的是，光子（玻色子）可无限叠加于同一量子态，因此光子计算具有支持大规模并行运算的天然优势。光子计算这种物理-算法的深度耦合巧妙地避开了“内存墙”瓶颈以及计算能耗等问题。依靠低延迟、低功耗、大带宽和先天并行计算架构等优势，光学神经网络正引发高性能计算的范式研究革新，成为新一代人工智能计算发展的国际前沿。

光子计算根据架构功能可以分为光学算子和光学计算系统。其中，光学算子通常为具有特定数学运算功能的光计算单元，可根据具体功能划分为微分算子、积分算子、卷积核算子、离散余弦变换算子等种类，具有易制备、轻量化、灵活性强的优点。光学计算系统根据网络架构和功能，可分为模拟伊辛模型解决组合优化问题的光子伊辛机、模拟矩阵乘加运算实现基本判

断功能的全连接光学神经网络、使用模拟卷积核功能的光学器件实现特征提取的卷积光神经网络等。光计算体系不断推陈出新，为后摩尔时代的算力突破注入创新活力。

1 特定功能的光子计算

1.1 光学算子

偏微分方程（partial differential equations, PDEs）是涉及一个或多个未知函数及其偏导数的方程，在通信、成像、传感、量子技术等领域被广泛应用。传统数值解法（如有限差分法）的计算复杂度高，难以满足高通量、低能耗、实时性的计算需求。与电子计算不同的是，光学算子利用光的物理特性执行特定的数学运算（如积分、微分、卷积等），具有计算速度快、能耗低、并行性高的优点，能够有效提升计算能效[11, 12]。光学算子适配多种灵活的应用场景，既可以独立完成数学运算，也可以充当光电异构集成计算系统中的前置数据处理器实现计算加速。举例来说，光学微分算子可进行独立的图像边缘检测、光学傅里叶变换；光学积分算子可以用于图像降噪、三维重建；光学卷积算子可用于图像隐含特征提取，加速计算机视觉领域数据处理；离散余弦算子能够对连续信号进行稀疏编码，提升数据处理速度和动态图像识别准确率。

近年来，有关光学算子的研究成果层出不穷。北京理工大学黄玲玲课题组[13]设计了一种基于六方晶格周期性硅圆柱阵列的介电超表面，通过激发电环形偶极子与磁环型偶极子共振的协同失谐效应，实现了宽带拉普拉斯微分操作，无需4F系统即可实现实时图像边缘检测。该课题组[14]利用空间微分截止超表面和自适应液体棱镜组成了光学微分元件，实现了高效的空间微分器，并应用于可调节视场边缘检测。浙江大学马云贵课题组[15]利用多层介电薄膜堆叠结构构建了集成微分算子，实现了对波长尺度的超紧凑、高带宽、实时性的高分辨率图像边缘检测。北京大学王兴军课题组-常林课题组[16]利用时间-波长展宽的方法在光频梳驱动的片上光处理单元中实现了光学卷积运算，实现了每平方毫米每秒1万亿次运算的超高计算密度。清华大学陈宏伟课题组[17]研制了基于光学掩膜版的无透镜光学卷积计算算子，显著降低了设备尺寸和系统复杂度。

然而，光学算子通常只能支持单一类型的运算，无法独立处理复杂任务。为了进一步释放光学计算的算力优势和能效比，光学神经网络逐渐走进研究人员的视野。

1.2 光学伊辛机

伊辛模型在数学上可以抽象为涉及多个变量的组合优化（combinatorial optimization, CO）问题，在车辆调度、算力调度、电力优化、通信网络优化、生物医药研发、大规模制造等复杂系统组合优化问题中具有极强的现实意义。伊辛模型最早由德国物理学家威廉·楞次（Wilhelm Lenz）和他的学生恩斯特·伊辛（Ernst Ising）提出，用于描述铁磁体内部的原子自旋状态及其与宏观磁矩的统计物理模型。该模型需要找到磁体自旋变量+1或-1的最稳定排列，使系统总能量最小。该问题的难点在于系统中的每一个变量均影响其他变量的状态，因此建立在冯·诺依曼结构上的传统计算机求解伊辛问题时面临指数级计算量爆炸。与之不同的是，光子伊辛机[18]利用光学参量构建伊辛哈密顿量，通过物理系统的内在优化特性重构算法。自旋变量的寻优过程不再依赖矩阵乘法运算，而是经由光的干涉、散射等物理过程自然完成。中国科学院大学李明课题组[18]利用微波相位表示伊辛自旋，并使用长距离光纤作为存储介质产生25600个稳定的伊辛自旋，在解一维、二维等伊辛模型时展现出了极低能耗。瑞典哥德堡大学Johan Akerman课题组[19]报道了基于自旋波的时间复用相干伊辛机，构建了微型化、低能耗的高性能计算平台。新加坡国立大学Aaron Danner课题组[20]围绕自由空间马赫-曾德干涉仪构建了激光分束-合束系统，配合光电调制器控制每束光的相位和幅度，利用光束合并时产生的干涉图案获取优化问题的数学解。该系统能够处理涉及1000个以上相互连接变量的组合优化问题，显著提升了伊辛问题的求解能力。华中科技大学张新亮课题组[21]提出了基于光电耦合振荡器的单片集成四磁子伊辛机，提升系统集成度的同时降低了系统功耗和收敛时间。

光子伊辛机求解NP-HARD组合优化问题能够显著提高计算能效，然而其架构本质是针对特定问题的特化网络。随着智能计算向着多场景、多模态快速发展，光学神经网络这一更加通用的网络架构开始快速崛起。

1.3 光学神经网络

1.3.1 空间互连式光学神经网络

空间互连式光神经网络被称作智能光计算的物理试验场，其相关研究能够追溯到20世纪70年代。1978年，斯坦福大学的Goodman等人[22]首次介绍了一种非相干的光学信息处理方法。该方法构建了光矢量矩阵乘法的计算模型，可用于执行离散傅里叶变换。1982年，Hopfield[23]提出一种单层反馈神经网络，其核心功能是联想记忆和组合优化能力。该网络理论清晰，结构简单，适合小规模存储。Hopfield网络的出现为机器学习计算范式带来了一次重大变革。1985年，Farhat等人[24]采用了无聚焦全息互连的方案，首次实现基于光学方法的Hopfield网络。在随后

的20年时间里，伴随着光子学以及集成光子学的研究发展，光学神经网络的网络类型、计算功能以及技术实现方案变得日益多样化和复杂化。根据系统结构所具有的特征，主要能够划分为空间互连式以及片上集成式[25]这两大类别。

空间互连式光学神经网络的网络架构和全连接人工神经网络的拓扑结构颇为相似，网络结构由输入层、隐藏层和输出层组成，并且每一层中的神经元都与相邻层神经元相连接。其中，隐藏层通常以瑞利-索末菲衍射理论为设计基础，由衍射光学元件、非线性光学材料、电控调制器等调制器件构成。调制器件中的最小调制单元被称为光学神经元，神经元之间通过光场传输相互连接。当包含待处理信息的输入光场进入隐藏层之后，其振幅、相位以及偏振等光场信息受到光学神经元的物理调制，实现矩阵乘加运算。经过N个隐藏层处理的运算结果最终在输出层呈现。衍射式神经网络的设计成本相对较低、实验平台易搭建、调制过程直观，是验证光学神经网络性能的理想方式。2018年，林星等人[26]借助三维打印的介质材料构建了级联的衍射光学元件，进而在太赫兹波段构建出了多层感知器，首次达成了可以对MNIST数据集展开内容识别的全光深度神经网络（如图1（a）所示）。这项工作充分展示了全光神经网络低功耗、高并行性以及高运算速度等诸多特点，也由此开启了衍射深度神经网络（deep diffractive neural network, D^2NN）相关的研究热潮。

然而，衍射神经网络的功能和衍射结构之间存在极为紧密的相关性，一旦衍射结构完成制造，其拓扑结构以及计算功能便会随之固定下来，无法进行参数调整或功能重建。另外，由于衍射神经网络的工作原理是建立在惠更斯-菲涅尔原理之上的波前调制机制，因此这种网络主要适用于相干光源，对于更为普遍的非相干光场景则难以有效应对。同时，由于受到衍射角物理规律的限制，全连接衍射神经网络的相邻衍射层需要满足特定的衍射条件，如波长（λ）与孔径（d）的比值大于衍射角（θ）的正弦值时（$\lambda/d > \sin\theta$）波动会显著发生衍射现象，这就导致系统的体积很难被压缩，网络的集成度也相对较低。除此之外，由于缺乏有效手段实现深度神经网络的非线性激活函数，衍射神经网络实际上是一个由多个衍射层所构成的矩阵线性乘加的级联系统。

为应对光学神经网络灵活性有限的问题，崔铁军团队[27]在2022年提出一种可编程的光学神经网络（如图1（b）所示）。该网络是由五层可编程超表面阵列构成的，借助FPGA电路对超表面进行统一调制。该方案可以实时操控电磁波的传播路径，并处理传统的深度学习任务。此项工作一方面提供了可重构光学神经网络的技术解决方案，另一方面也为网络参数在线训练提供了相应的研究平台。2023年，Ozcan团队[28]打破了光学神经网络只能处理相干光的思维定式。

如图1（c）所示，研究团队使用大量的空间变化的相干点扩散函数，对非相干光场景下输入光场与输出光场的线性关系加以近似，证明了在空间非相干光的照明条件下，纯相位衍射网络是可以开展任意光强线性变换操作的。2025年，顾敏团队[29]利用双光子聚合打印技术，把边长为150 μm的衍射器件集成在多模光纤的端面上，实现了信号传输与信号解析高度集成的光计算系统（如图1（d）所示）。

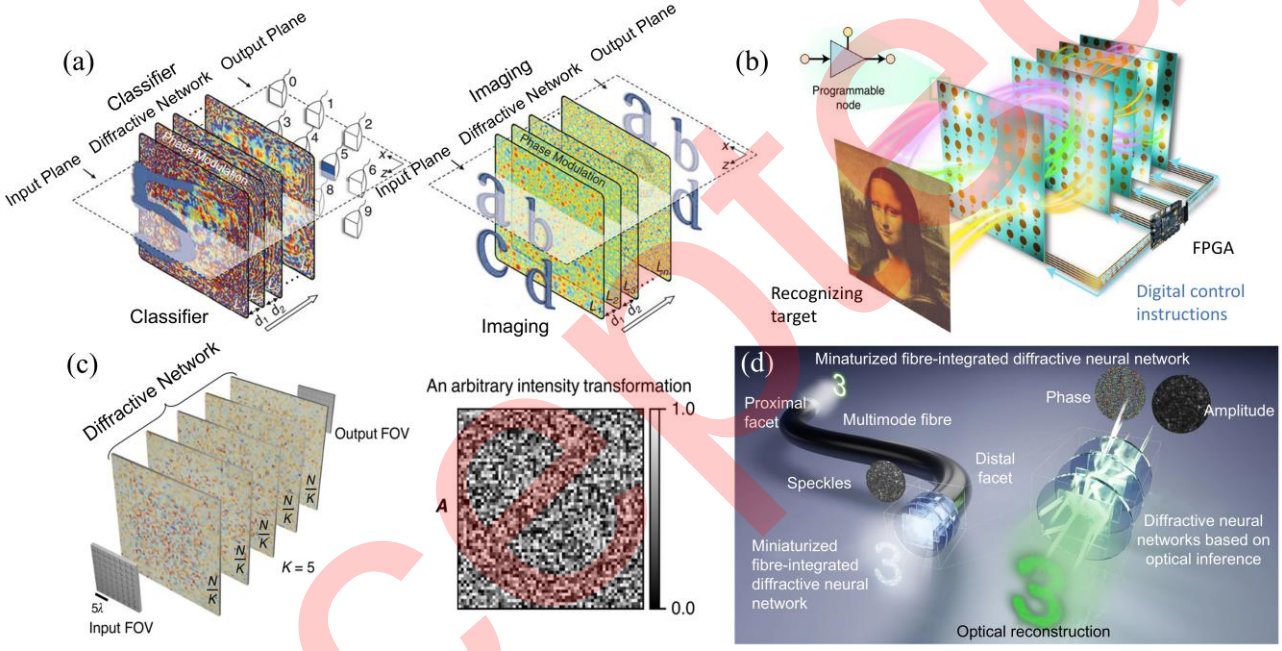

图1　空间互连式光学神经网络. (a) 基于光学衍射的自由空间全光机器学习网络框架[26]；(b) 基于FPGA的可编程衍射深度神经网络[27]；(c) 非相干光照明衍射深度神经网络[28]；(d) 光纤端面集成的衍射式神经网络[29]

Figure 1　Diffractive optical neural networks. (a) Free-space all-optical machine learning network framework based on optical diffraction[26]; (b) FPGA-programmable diffractive deep neural network[27]; (c) diffractive deep neural network with incoherent light illumination[28]; (d) fiber-end integrated diffractive neural network[29]

空间互连式光神经网络在系统稳定性、集成度和实用性方面存在一定局限性，但是其高通量计算能力以及对复杂视觉任务的高效处理能力，使其有望在轻量化边缘识别等应用场景中开辟特殊赛道。近几年基于D^2NN架构的光学神经网络研究持续涌现，主要包含两方面研究趋势：一方面，针对光场调控、计算成像等领域持续拓展应用边界；另一方面，则聚焦于实现全光非线性激活函数以增加网络的深度和泛化能力。总体而言，若要将空间互连式光学神经网络更好地推向产业化应用当中，不可避免地要解决如下几个问题。

（1）空间互连式光神经网络的系统集成

空间互连式光学神经网络主要通过超表面、相位掩模版、空间光调制器等衍射光学器件构建计算系统，通常受到系统尺寸庞大、安装精度要求高的限制。提升系统集成度以及与现有光计算平台异构集成，成为了空间互连式光学神经网络迈向产业应用的重要一步。Cui等人[30]在图像光谱仪前端集成全光卷积层等预处理模块，显著增强了系统的信息处理能力；Ren等人[31]在边缘计算设备的图像处理前端搭载全光离散余弦变换通用计算模块，实现信息的高效降维处理。一系列基于空间互连式光学神经网络的创新应用展示出了卓越的性能表现，但同时也对微纳加工的精度和良品率控制提出了极高的要求。可见，空间互连式架构在实现大规模产业化应用的技术路径中，系统集成问题不容忽视。

（2）非线性激活函数功能的实现

空间互连式光学神经网络利用光波的衍射行为构建了独特的计算机制，在全连接层、卷积层等光学神经网络的核心设计架构中提供了万亿次运算（tera operations per second, TOPS）量级的超高算力指标。但是在其计算过程中仅包含光场调制以及衍射乘加过程这两类线性计算，这一特性导致空间互连式光学神经网络难以实现深度计算，网络有效层数等效为一层。另外，非线性激活函数的缺失同时限制了网络对复杂函数的拟合能力以及高阶特征提取能力。针对这一关键问题，研究人员也探索出多种技术路径：如通过在衍射层前端引入铁电薄膜[32]、二维材料[33, 34]等具有显著非线性响应特性的功能材料，来构建ReLu、Sigmoid、Tanh等经典非线性激活函数的关系映射；采用饱和吸收体[35]、铌酸锂[36]等材料模拟非线性传输特性；利用线性结构光场的迭代输入机制，诱导产生独特的光学非线性效应等，然而目前仍缺乏通用型、低功耗、快速响应的非线性激活函数实现手段。

（3）网络权重参数的编码与重构

空间互连式光学神经网络的参数由超表面、相位掩膜版的结构决定，权重参数在衍射元件加工完毕后无法灵活变更，网络不具备可编程性。采取FPGA编码多层超表面阵列[27]实现动态编程调控，或设计可插拔[37, 38]的衍射模块均可以一定程度上切换网络功能，然而这种物理切换网络功能的方式在通用性和便捷性上有所欠缺。空间光调制器（spatial light modulator, SLM）和数字微镜阵列（digital micromirror device, DMD）等先进光学器件已能实现调制表面的刷新率超100帧每秒（frames per second, fps）的动态调控，但这类基于液晶分子等阵列的技术仅能支持对特定入射光的相位调制，难以充分释放光神经网络的应用潜力。

1.3.2 片上集成式光学神经网络

片上集成式光学神经网络在集成度、能效比，以及制备工艺、成本控制方面具有显著优势，是一种极具应用潜力的技术方案。依据片上集成光学单元的不同类别，大致能够划分为基于马赫-曾德干涉仪（Mach-Zehnder interfermetor，MZI）、微环谐振腔（micro ring resonator, MRR）、亚波长衍射结构及其他类型的片上集成光学神经网络。其基本的工作原理在于，通过片上光学器件对信号光的干涉与衍射，进而将矩阵运算转变为光场调制的物理过程。矩阵经过奇异值分解之后，可以被分解为一个对角矩阵和两个酉矩阵，其中对角矩阵可通过可调光衰减器实现，而$N\times N$酉矩阵则可以通过多个光学分束器和移相器的串联来实现。以MZI为例，单个MZI包含了两个分束器以及两个移相器，因此MZI可以同步执行分束以及移相任务。2017年，麻省理工学院沈亦晨团队[39]在硅光子平台上集成了56个可编程的MZI单元（如图2（a）所示），凭借MZI级联的方式构建起了全连接神经网络，并顺利完成了4×4维度的光学矩阵乘积运算。他们对网络针对元音数据集的识别能力展开了实验测试，以此验证了片上集成光学神经网络的计算性能。2021年，Gu Mile团队[40]着手开发了一种能够对相位和幅度信息加以编码的ONN系统平台，该平台能够通过光的干涉来执行复数运算，从而有效提升网络的计算速度以及计算能效（如图2（b）所示）。2022年，蒋旭东团队[41]在MZI级联阵列当中引入了衍射单元，用于实现傅里叶变换及逆变换。该方案在提高网络的可处理矩阵维度的同时，降低了计算能耗（如图2（c）所示）。2023年，祝宁华团队[42]借助两个4×4多模干涉耦合器以及四个移相器，构建出了三个2×2相关的实值卷积核，进而研制了超高集成度的光学卷积处理器，成功完成了卷积运算当中的加法操作以及卷积核的动态重构（如图2（d）所示）。2025年，清华大学陈宏伟课题组[43]构架了基于存内光计算的多任务处理网络架构，将片上光学的参数存储进固定的光学元件，利用电子元件对少数参数进行调节。该工作利用深度回归算法对光的物理传播模型进行建模，构建了紧凑的衍射芯片。张新亮团队在热光移相器阵列集成的片上平台中，配合锗激活函数的辅助和衍射结构的波长相关性，构建了可以多线程并行处理的全光非线性衍射深度神经网络芯片，这种新颖的架构在集成密度、多线程推理、输入大小的可扩展性以及网络深度方面均有显著优势[44]。

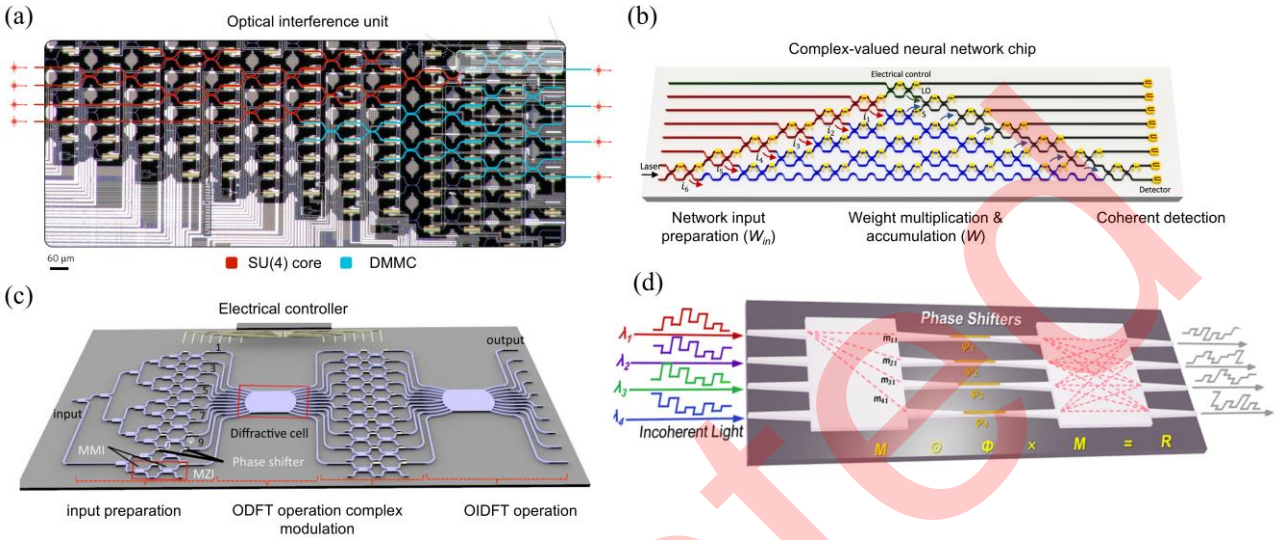

图2　片上集成式光学神经网络。 (a) 基于MZI级联阵列的全光前馈神经网络架构[39]; (b) 可实现复数运算的光学神经网络[40]; (c) 基于MZI级联阵列和衍射单元的集成光学神经网络[41]; (d) 基于多模干涉耦合器和相移器的光学卷积处理器[42]

Figure 2　On-chip integrated photonic neural networks. (a) All-optical feedforward neural network architecture based on Mach–Zehnder interferometer (MZI) cascaded arrays[39]; (b) optical neural network enabling complex-valued operations[40]; (c) Integrated photonic neural network based on MZI cascaded arrays and diffractive units[41]; (d) optical convolutional processor utilizing multimode interference couplers and phase shifters[42]

片上集成式光学神经网络得益于聚焦离子束刻蚀（focused ion beam, FIB）、电子束光刻（electron beam lithography, EBL）等片上加工工艺的技术沉淀，在系统集成度、器件精度等方面拥有明显优势。然而，基于片上集成工艺的光学芯片在产业化进程中面临参数训练过程复杂、网络规模难以扩展的问题。

（1） 片上集成光学神经网络的参数调制

片上集成式光神经网络可以利用热光相位调制、相变材料调制等方式实时调节网络参数，构成可编程的光计算网络。然而受到调制机制、集成工艺、插入损耗等因素的限制，调制效率与光计算效率在能耗和响应速度方面严重失配。举例来说，MZI阵列构成的光子芯片利用硅材料的光学特性（如热光效应）实现对光波相位和强度的电控调制，需要集成微型发热电极对每一个MZI做独立的相位调控以实现矩阵运算。热扩散时间常数约为毫秒量级，调制带宽不超过1千赫兹（kilohertz, kHz）。同时，单个移相器的功耗可达毫瓦（milliwatt, mW）量级，极大影响了光计算的能效优势，也限制了网络的应用范围。

（2） 片上集成光学神经网络的网络规模

另外，片上集成光学神经网络的神经元数量取决于片上集成光子器件的数目。以MZI阵列集成方案为例，单个计算单元尺寸在数十至数百微米，单芯片神经元数量通常不超过10^3个。此外，光波导之间的串扰、热噪声和调制系统复杂度将随着神经元数目增加而加剧[45]。针对网络规模提升困难的问题，学界展开了一系列研究和讨论，如提升亚波长微纳结构单元的制备工艺；引入相变材料实现非易失性光调谐，降低持续热调谐需求[46]；利用三维集成封装的技术来隔离热敏感元件；开发新型低损耗材料（如铌酸锂薄膜集成）或拓扑光子结构来抑制模式串扰或者探索动态热补偿算法，以实时校准温度引起的相位漂移等。如何以可控成本实现网络规模提升，是片上集成光学神经网络迈向实用化和产业化的关键。

表1　部分典型光学神经网络计算架构、网络能效和应用对比

Table 1　Comparision with optical neural networks in energy efficiency and application scenarios

来源	单位面积算力	计算能效比	神经元总数	可调神经元总数	网络规模	功能应用及特点
Shen等人[47]	不适用	不适用	213	213	1065	4分类-元音识别
Feldmann 等人[48]	162.00 TOPS/mm^2	0.50 TOPS/W	64	64	29186	10分类-手写数字识别
Zhou等人[49]	不适用	0.71 TOPS/W	49万	49万	147万	10分类-手写数字识别
Ashtiani 等人[50]	1.75 TOPS/mm^2	2.90 TOPS/W	67	67	67	4分类-字母字符识别
Xu等人[51]	不适用	不适用	867	867	不适用	10分类-手写数字识别
Wang等人[52]	30000 TOPS/mm^2 (10 G调制器)	不适用	900	不适用	不适用	图像分类
Yan等人[53]	130 TOPS/mm^2 (30 Gb调制器)	8.26 POPS/W (30 Gb调制器)	不适用	不适用	不适用	图表示学习 半监督学习，动作识别
Fu等人[54]	4600 TOPS/mm^2 (10 G调制器)	90900 TOPS/W (10 G调制器)	372	不适用	不适用	各种任务的多通道分类
太极芯片[55]	1758 TOPS/mm^2	160.82 TOPS/W	4256	160	1396万	1000种分类，多功能内容生成
Liu等人[43]	132.1 TOPS/mm^2	1.12 TOPS/W (703.5 TOPS/W mm^2)	100	100	100	多类别任务判断及回归问题
英伟达 Tesla V100[56]	0.037 TOPS/mm^2	0.1 TOPS/W	不适用	不适用	不适用	通用计算机科学
英伟达 H100 PCIE[57, 58]	不适用	0.15 TOPS/W	不适用	不适用	不适用	通用计算机科学

a) NVIDIA系列产品计算能力采用FP64 Tensor Core精度的峰值进行计算

2 光学神经网络的破局与展望

光学神经网络在矩阵乘加、傅里叶变换等核心计算功能方面所呈现出的算力指标，相较于电学处理器而言是更为出色的。然而时至今日，光学神经网络依旧没有完全走出实验室验证阶段。除去前文提到的空间互连式与片上集成式光学神经网络计算架构自身的限制因素，以及微纳光子器件的工艺复杂度高、集成难度大等工程方面的因素外，光学神经网络若想真正实现产业化落地还必须突破模型训练以及推理计算这两个极为关键的技术环节。

2.1 光学神经网络的训练

当下，光学神经网络的权重参数训练大多采用离线训练模式，其具体实现流程可大致分为三个环节：其一，建立神经网络的数学模型，并且构建损失函数；其二，借助反向传播算法对损失函数的梯度进行计算，而后根据梯度下降法逐步对权重参数予以优化；其三，在模型结果完成收敛后，依照权重参数设计并制备与之对应的物理结构。这一过程往往需要反复开展多次，以此来缩小计算机离线训练所得到的模型与实际光学系统之间存在的差异。然而，因为环境噪声、机械抖动以及装配误差等诸多因素的影响，离线训练所得到的网络性能和实际部署之间往往存在一定的偏差。伴随着网络规模的不断增大，离线训练以及在线部署所需要的计算资源、误差补偿以及调试成本都会呈现指数级的增长态势，这无疑对光学神经网络的规模以及工作效率形成了严重的制约。

为了消除数值仿真与光学系统之间存在差异而引发的误差问题，发展在线（原位）学习机制[59]十分必要。这一认知推动了非反向传播的优化算法研究。2018年，Hughes等人[60]依据MZI级联的方式提出一种原位训练方法，能够直接在硬件层面完成对神经网络参数的训练，加快了光学神经网络的参数迭代并提升了推理效率（如图3（a）所示）。2020年，Zhou等人[61]提出了一种在线训练衍射神经网络的方法，借助级联的空间光调制器把反向传播算法以光学形式应用于线性衍射神经网络当中，最终获得了与离线训练准确率相当的分类准确率（如图3（b）所示）。2021年，Zhou等人[62]提出一种利用可编程光电器件来构建大规模且复杂的神经网络的可重构衍射处理单元DPU（（如图3（c）所示）），通过自适应训练算法直接在器件上开展物理求导以及物理结构参数更新等操作，加快了网络参数的迭代进程。2024年，Dirk England课题组[63]采用随机方向扰动法（如图3（d）所示）成功实现了不依赖反向传播的高速原位训练目标。Zheng等人[64]提出了双边训练方法（如图3（e）所示），将输入信息经过光学以及数字编码后分别输入到物理模型和数学模型中，在物理模型中构建误差预测网络，然后凭借数值模型获取网络内部的状态，由此展开参数的迭代更新。

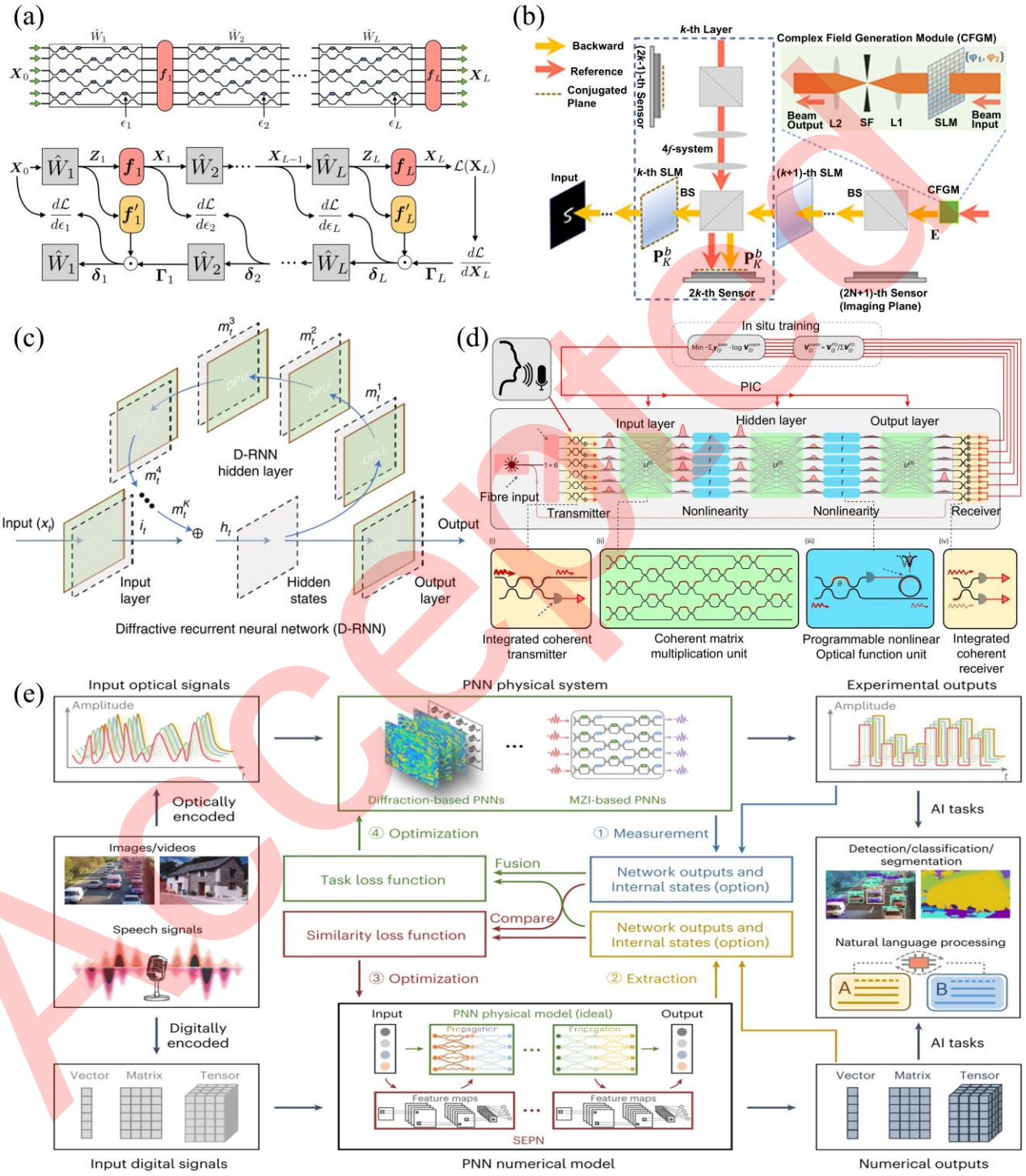

图3　光学神经网络的在线训练方法. (a) 基于原位误差反向传播的片上集成式光学神经网络在线训练方法[60]; (b) 空间衍射式光学神经网络在线训练方法[61]; (c) 基于可编程衍射处理单元的大规模神经形态光子计算[62]; (d) 基于随机方向扰动的前向光学神经网络训练方法[63]; (e) 基于双边自适应的光学神经网络训练[64]

Figure 3 Online training methods for optical neural networks. (a) On-chip integrated optical neural network training methods based on *in situ* error backpropagation[60]; (b) online training method for spatial diffractive optical neural network[61]; (c) large-scale neuromorphic photonic computing based on programmable diffractive processing units[62]; (d) forward-only training method for single-chip photonic deep neural network[63]; (e) dual adaptive training of photonic neural networks[64]

梯度下降算法在面对大规模复杂模型时难以做到良好匹配，而基于扰动的优化方法在效率方面存在瓶颈。针对光学神经网络训练的技术难题，清华大学方璐、戴琼海团队[65]在2024年提出了“空间互易-时间反演”的对称光学传播模型（如图4所示）。该研究团队指出，鉴于物理学所具有的对称性特点，反向传播算法中的梯度数据是能够从光场的前向传播过程中通过测量方式获取到的。因为不需要进行反向传输操作，所以光的传输路径可以始终保持一致，天然地避开了由于传输路径出现不对齐情况而引发的计算误差问题。这种全前向的智能光计算在线训练架构，从其根本上便对物理计算的精确性给予了有力保障。这项工作使得智能光计算网络成功摆脱了对GPU离线训练的依赖，同时也为构建高效且精确的光学智能训练系统开辟出了一条全新的技术路径。

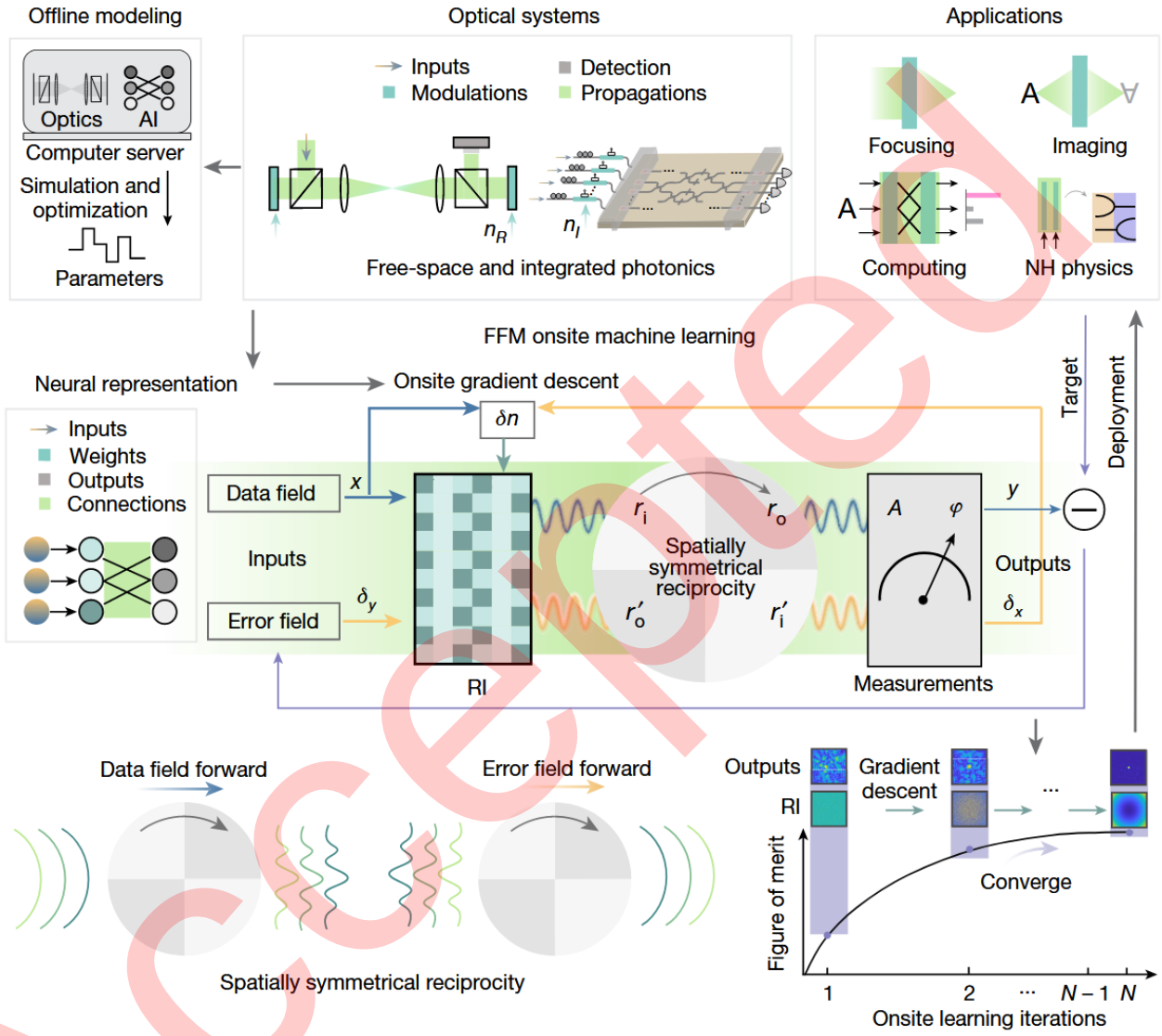

图4 “空间互易-时间反演”的光学传播模型及全前向智能光计算的在线训练架构[65]

Figure 4 “Spatial symmetrical reciprocity” optical propagation model and an online training method for fully forward optical neural networks[65]

2.2 光学神经网络的推理

尽管光学神经网络拥有高速传输、高维信息并行处理以及计算过程低能耗等诸多优点，然而受到光学器件规模以及器件调制难度方面的限制，它的真实算力通常低于通用的电学神经网络。就可训练参数而言，电子神经网络能达到百万量级（即10^6），而当下主流的集成光子芯片仅仅能实现数百（即10^2）量级的参数调控。所以在现阶段，光学神经网络常常被用于简单的图像分类（如MNIST手写数字识别、CIFAR字符分类）以及边缘识别这类轻量化的数据处理任务中。

为了实现大规模智能光子计算，单纯靠扩大光学元件规模或者提升器件集成度一类的做法实际上并不现实，因为系统中的模拟噪声以及调制误差会随着器件规模的上升而呈现几何级数般的放大态势。正是由于这种“规模不经济”状况的存在，使得光子计算的实际算力与理论所展现的性能之间出现了明显的差距。针对大规模智能光计算所面临的架构难题，清华大学方璐、戴琼海团队[55]另辟蹊径地提出了一种基于分布式广度的光计算理论框架“太极”。如图5所示，“太极”芯片给出了一种与传统深度神经网络不同的计算框架，它将空间衍射和片上干涉两种光计算模式结合起来，并引入纠错输出编码（error-correcting output codes, ECOC）机制构建多通道决策体系，从而形成了一套高效的用于任务分配的协议。具体来讲：该框架会借助基于ImageNet数据集预训练得到的固定平面衍射波导对输入网络的高维信息进行特征提取，将信息由64通道压缩至8通道后输入进由可编程MZI阵列所构成的干涉计算单元当中，利用奇异值分解进行矩阵运算。由该架构的衍射结构负责对输入数据的通用基础特征（如边缘、纹理、方向等）进行提取，同时滤除统计上的冗余信息。衍射模块在整个计算系统中承担感知前端的职能，负责对海量的原始数据进行快速预处理，生成一个信息密度极高的紧凑表征。而电调控的可重构干涉计算单元则实现了双重功能突破：一方面，干涉单元精准对接前端衍射计算模块输出的“特征摘要”，通过动态参数调整对特征进行深度、有针对性的分析，以满足多样化任务的计算需求；另一方面，MZI阵列单元具有可重构特性以及特异性调整能力，为构建大规模、复杂判断策略的光学计算系统提供了核心支撑——使其不仅能够有能力高效执行ECOC方案分配的特定子任务，还能依据具体计算的权重需求动态调整计算路径，以此实现片上光计算网络的多通道协同判断运算功能。

实验结果表明，太极芯片在多个基准数据集的性能测试中表现亮眼。值得强调的是，太极芯片还采用了“编码-解码”的闭环处理架构：在干涉计算单元的输出端集成用于解码的衍射计算单元，与输入端的编码器形成了操作互补的信息恢复机制。这一设计既确保了系统的层次化深度运算，又保证了架构计算过程中数据表征的一致性，为后续数据的可读性处理预留了兼容接口，同时还为未来系统级的应用拓展提供了可能。

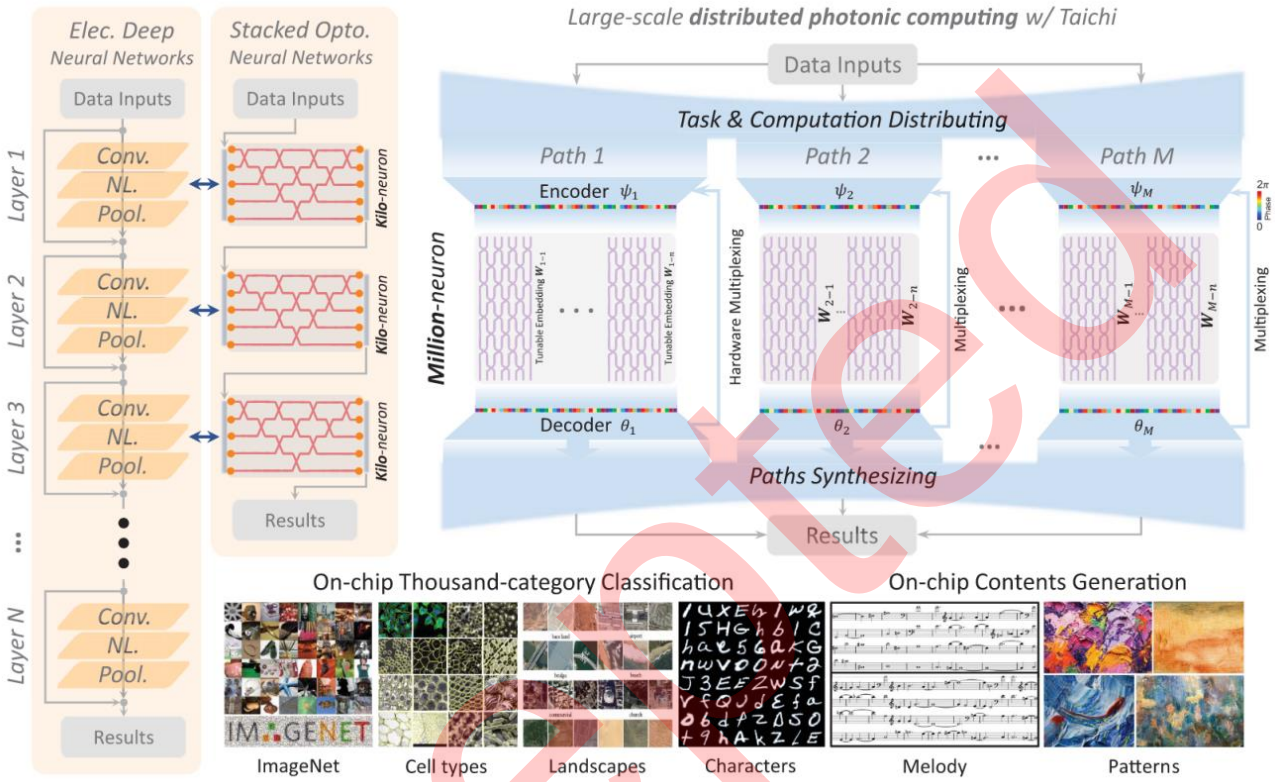

图5　基于干涉-衍射联合传播模型的大规模通用智能光计算芯片“太极”架构及功能[55]

Figure 5　Architecture and functions of the large-scale photonic computing chiplet “taichi” based on integrated diffractive-interference hybrid design[55]

研究团队运用这种“广度优先”的计算架构，设计并制造出了国际首款大规模通用智能光计算芯片“太极”，在确保系统结构紧凑性的前提下，达成了对高维信息的高效处理。此项工作探索了光学神经网络在复杂认知任务（如自然场景千类对象识别、跨模态内容生成）中的实际应用尝试，为智能光子计算走向产业化发展提供了极为重要的技术支撑。在系统性能方面，“太极”芯片达成了160 TOPS/W（每秒万亿运算每瓦特）的系统能效，并且其制备工艺的要求仅为百纳米级别。这项工作构建起了完整的光子计算技术链条，赋予其能够支持通用人工智能（artificial general intelligence, AGI）大模型训练以及推理的能力。据此，光计算芯片有希望凭借全新的计算范式来破解当下人工智能所面临的算力困境，以更少的资源消耗以及更低的边际成本，给人工智能大模型训练、通用人工智能的发展以及复杂智能系统的构建提供高速、高效并且节能的创新解决办法。

3 结语

随着人工智能向多模态感知、实时决策与复杂认知等方面快速演进，“后摩尔时代”对算力的需求呈现出指数级增长的态势。光计算技术依靠光子传播的天然并行性以及超低能耗特性，在实时图像重建、高维信号处理以及多模态大模型推理等场景里展现出了巨大潜力，成为突破算力困境的关键技术路径。近年来，为了满足日益增长的复杂推理任务对计算效率和精度的严格要求，光计算系统正加速朝向高集成度、高能效比以及异构化集成的方向发展。

不过，在光计算系统向通用化、规模化发展的过程中，系统在稳定性、集成度及实用化水平等方面仍存在显著的技术瓶颈。具体而言，其物理架构的脆弱性可能导致环境适应性不足，高密度集成工艺面临微纳加工精度与良率控制的挑战，而现有系统在实时动态响应和跨平台兼容性方面的局限性也制约了其实际部署能力。除物理因素和工程化制约问题外，光计算系统的架构设计方面仍大有可为。从神经网络的功能来看，光子计算天然适合进行矩阵乘法一类的线性运算，但非线性激活函数难以高效实现。因此，多数光学神经网络仅为浅层的线性模型，难以处理复杂任务。现有的非线性运算方案依赖电学系统辅助或者全光非线性材料，存在响应速度慢、能耗高或者集成度低的问题。从网络的训练机制来看，由于光子状态无法存储，光计算无法通过自身系统完成梯度反向传播和权重更新，需要离线训练或借助光电转换进行辅助计算，导致系统延迟较高。从搭载网络的硬件角度来看，现有光计算架构（如衍射式神经网络）的结构决定了网络功能，任务切换需要重新设计光学元件。光场维度复用可以实现多任务并行，但是算法设计难度会显著上升，需要根据光传播特性定制网络结构。此外，构建光计算异构计算平台时，还需要突破系统集成度、运行稳定性以及可扩展性等方面的技术瓶颈。

不可否认，太极芯片的问世让世界看到了光子计算架构迈向产业化应用的可能性，标志着光子计算架构向产业化迈出了实质性的步伐。该芯片不仅成功实现了大规模高难度图像分类、多功能内容（包括音乐创作、风格化图像合成等）生成等核心功能，更在上述提到的与全前向传播结合的光学计算系统的协同研究中，展现了在突破网络深度极限、散射介质成像、非视域成像等领域的应用潜力。这些突破性进展不仅为光子计算相关的研究提供了可验证的技术路径，更确立了产业化的技术标杆。总结来说，光子计算的未来发展将

遵循专用化优先，异构融合渐进的技术路径。从短期来看，光子计算适合被用于处理通用化、轻量化的任务，如图像处理、光通信等并行度高、精度容忍度高的场景。这类任务可以有效发挥出现阶段光子计算系统的并行计算能力以及低能耗的优势。从长期发展来看，研究光神经网络在线训练、光电异构集成、算法-硬件协同优化，开发适配光传播物理规律的原生算法，释放光计算多通道处理潜力，逐步扩展应用边界更为重要。值得注意的是，若要推动该技术走向产业化应用，除了需突破上述核心瓶颈外，还需考虑如何平衡系统性能、制造成本与商业化落地之间的矛盾，以及建立适配不同应用场景的标准化评估体系。这些问题的协同解决将决定以光学神经网络为主的光计算系统能否真正实现从实验室创新到产业落地的跨越。伴随着技术成熟度的提升以及应用场景的拓展，光计算有望成为新一代计算架构中占据核心地位的技术引擎。

参考文献

[1] Moore GE. Cramming more components onto integrated circuits. Proceedings of the IEEE, 1998, 86: 82-85
[2] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436-444
[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, Dec 04-09, 2017
[4] Yang Q, Luo Z-D, Duan H, et al. Steep-slope vertical-transport transistors built from sub-5 nm thin van der waals heterostructures. Nature Communications, 2024, 15: 1138
[5] Fu T, Zhang J, Sun R, et al. Optical neural networks: Progress and challenges. Light: Science & Applications, 2024, 13: 263
[6] Zhang D, Tan Z. A review of optical neural networks. Applied Sciences, 2022, 12: 70773-70783
[7] Wetzstein G, Ozcan A, Gigan S, et al. Inference in artificial intelligence with deep optics and photonics. Nature, 2020, 588: 39-47
[8] Liao K, Dai T, Yan Q, et al. Integrated photonic neural networks: Opportunities and. Acs Photonics, 2023, 10: 2001-2010
[9] Hu J, Mengu D, Tzarouchis DC, et al. Diffractive optical computing in free space. Nature Communications, 2024, 15: 1525
[10] He C, Shen Y, Forbes A. Towards higher-dimensional structured light. Light: Science & Applications, 2022, 11: 205
[11] Tang Y, Chen R, Lou M, et al. Optical neural engine for solving scientific partial differential equations. Nature Communications, 2025, 16: 4603
[12] Yuan H, Du Z, Qi H, et al. Microcomb-driven photonic chip for solving partial differential equations. Advanced Photonics, 2025, 7: 016007
[13] Zhou C, Muhammad N, Zhao R, et al. Metasurface enabled broadband, high numerical aperture laplace differentiator under multiple polarization illumination. Photonix, 2025, 6: 10
[14] Zhou Y, Li L, Zhang J, et al. Meta-device for field-of-view tunability via adaptive optical spatial differentiation. Advanced Science, 2025, 12: 2412794
[15] Zhou Y, Zhan J, Chen R, et al. Analogue optical spatiotemporal differentiator. Advanced Optical Materials, 2021, 9: 2002088
[16] Bai B, Yang Q, Shu H, et al. Microcomb-based integrated photonic processing unit. Nature Communications, 2023, 14: 66
[17] Shi W, Huang Z, Huang H, et al. Loen: Lensless opto-electronic neural network empowered machine vision. Light: Science & Applications, 2022, 11: 121
[18] Cen Q, Ding H, Hao T, et al. Large-scale coherent ising machine based on optoelectronic parametric oscillator. Light: Science & Applications, 2022, 11: 333
[19] Litvinenko A, Khymyn R, González VH, et al. A spinwave ising machine. Communications Physics, 2023, 6: 227-238

[20] Gao Y, Qi L, Lin H-L, et al. All-optical interferometer-based ising machine. Optica, 2025, 12: 831-840
[21] Wu B, Zhang W, Zhang S, et al. A monolithically integrated optical ising machine. Nature Communications, 2025, 16: 4296
[22] Goodman JW, Dias AR, Woody LM. Fully parallel, high-speed, incoherent optical method for method for performing discrete fourier-transforms. Optics Letters, 1978, 2: 1-3
[23] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 1982, 79: 2554-2558
[24] Farhat NH, Psaltis D, Prata A, et al. Optical implementation of the hopfield model. Applied Optics, 1985, 24: 1469-1475
[25] Zhou H, Dong J, Cheng J, et al. Photonic matrix multiplication lights up photonic accelerator and beyond. Light: Science & Applications, 2022, 11: 30
[26] Lin X, Rivenson Y, Yardimei NT, et al. All-optical machine learning using diffractive deep neural networks. Science, 2018, 361: 1004-1008
[27] Liu C, Ma Q, Luo ZJ, et al. A programmable diffractive deep neural network based on a digital-coding metasurface array. Nature Electronics, 2022, 5: 113-122
[28] Rahman MSS, Yang X, Li J, et al. Universal linear intensity transformations using spatially incoherent diffractive processors. Light: Science & Applications, 2023, 12: 195
[29] Yu H, Huang Z, Lamon S, et al. All-optical image transportation through a multimode fibre using a miniaturized diffractive neural network on the distal facet. Nature Photonics, 2025, 19: 486-493
[30] Cui K, Rao S, Xu S, et al. Spectral convolutional neural network chip for in-sensor edge computing of incoherent natural light. Nature Communications, 2025, 16: 81
[31] Ren H, Feng Y, Zhou S, et al. All-optical dct encoding and information compression based on diffraction neural network. Acs Photonics, 2025, 12: 1196-1211
[32] Yan T, Wu J, Zhou T, et al. Fourier-space diffractive deep neural network. Physical Review Letters, 2019, 123: 023901
[33] Tong L, Bi Y, Wang Y, et al. Programmable nonlinear optical neuromorphic computing with bare 2d material mos2. Nature Communications, 2024, 15: 10290
[34] Chen C, Yang Z, Wang T, et al. Ultra-broadband all-optical nonlinear activation function enabled by mote2/optical waveguide integrated devices. Nature Communications, 2024, 15: 9047
[35] Spall J, Guo X, Lvovsky AI. Training neural networks with end-to-end optical backpropagation. Advanced Photonics, 2025, 7: 016004
[36] Yildirim M, Oguz I, Kaufmann F, et al. Nonlinear optical feature generator for machine learning. APL Photonics, 2023, 8: 106104
[37] He C, Zhao D, Fan F, et al. Pluggable multitask diffractive neural networks based on cascaded metasurfaces. Opto-Electronic Advances, 2024, 7: 230005
[38] Wang G, Zang X, Tan Z, et al. Modular diffractive neural networks using cascaded metasurfaces. Laser & Photonics Reviews, 2025, e00923
[39] Shen Y, Harris NC, Skirlo S, et al. Deep learning with coherent nanophotonic circuits. Nature Photonics, 2017, 11: 441-447
[40] Zhang H, Gu M, Jiang XD, et al. An optical neural chip for implementing complex-valued neural network. Nature Communications, 2021, 12: 457
[41] Zhu HH, Zou J, Zhang H, et al. Space-efficient optical computing with an integrated chip diffractive neural network. Nature Communications, 2022, 13: 1044
[42] Meng X, Zhang G, Shi N, et al. Compact optical convolution processing unit based on multimode interference. Nature Communications, 2023, 14: 3000
[43] Liu W, Huang Y, Sun R, et al. Ultra-compact multi-task processor based on in-memory optical computing. Light: Science & Applications, 2025, 14: 134
[44] Zhang J, Wu B, Zhang S, et al. Highly integrated all-optical nonlinear deep neural network for multi-thread processing. Advanced Photonics, 2025, 7: 046003
[45] Lin J, Yang K, Fu Q, et al. A robust mzi-based optical neural network using qr decomposition. Journal of Lightwave Technology, 2025, 43: 1024-1031
[46] Zhang S, Wang W, Shen Z, et al. On-chip non-volatile reconfigurable phase change topological photonics. Advanced Materials, 2025, 37: 2418510
[47] Shen YC, Harris NC, Skirlo S, et al. Deep learning with coherent nanophotonic circuits. Nature Photonics, 2017, 11: 441-+
[48] Feldmann J, Youngblood N, Karpov M, et al. Parallel convolutional processing using an integrated photonic tensor core. Nature, 2021, 589: 52-58
[49] Zhou TK, Lin X, Wu JM, et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nature Photonics, 2021, 15: 367-373
[50] Ashtiani F, Geers AJ, Aflatouni F. An on-chip photonic deep neural network for image classification. Nature, 2022, 606: 501-506
[51] Xu X, Tan M, Corcoran B, et al. 11 tops photonic convolutional accelerator for optical neural networks. Nature, 2021, 589: 44-51
[52] Wang Z, Chang L, Wang F, et al. Integrated photonic metasystem for image classifications at telecommunication wavelength. Nature Communications, 2022, 13: 2131

[53] Yan T, Yang R, Zheng Z, et al. All-optical graph representation learning using integrated diffractive photonic computing units. Science Advances, 2022, 8: eabn7630
[54] Fu T, Zang Y, Huang Y, et al. Photonic machine learning with on-chip diffractive optics. Nature Communications, 2023, 14: 70
[55] Xu Z, Zhou T, Ma M, et al. Large-scale photonic chiplet taichi empowers 160-tops/w artificial general intelligence. Science, 2024, 384: 202-209
[56] Yao P, Wu HQ, Gao B, et al. Fully hardware-implemented memristor convolutional neural network. Nature, 2020, 577: 641-646
[57] Nvidia h100 tensor core gpu architecture. https://resourcesnvidiacom/en-us-hopper-architecture/nvidia-h100-tensor-c
[58] Nvidia h100 tensor core gpu datasheet. https://resourcesnvidiacom/en-us-hopper-architecture/nvidia-tensor-core-gpu-datasheet
[59] Buckley SM, Tait AN, McCaughan AN, et al. Photonic online learning: A perspective. Nanophotonics, 2023, 12: 833-845
[60] Hughes TW, Minkov M, Shi Y, et al. Training of photonic neural networks through in situ backpropagation and gradient measurement. Optica, 2018, 5: 864-871
[61] Zhou TK, Fang L, Yan T, et al. In situ optical backpropagation training of diffractive optical neural networks. Photonics Research, 2020, 8: 940-953
[62] Zhou T, Lin X, Wu J, et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nature Photonics, 2021, 15: 367-373
[63] Bandyopadhyay S, Sludds A, Krastanov S, et al. Single-chip photonic deep neural network with forward-only training. Nature Photonics, 2024, 18: 1335-1343
[64] Zheng Z, Duan Z, Chen H, et al. Dual adaptive training of photonic neural networks. Nature Machine Intelligence, 2023, 5: 1119-1129
[65] Xue Z, Zhou T, Xu Z, et al. Fully forward mode training for optical neural networks. Nature, 2024, 632: 280-286

Breakthroughs in Large-Scale Intelligent Photonic Computing Inference

Jing Pan[1†], Zhihao Li[1†] & Min Qiu[1,2*]

[1] Zhejiang Key Laboratory of 3D Micro/Nano Fabrication and Characterization, Department of Electronic and Information Engineering, School of Engineering, Westlake University, Hangzhou 310030, China
[2] Westlake Institute for Optoelectronics, Hangzhou 311421, China
† Equally contributed to this work
* Corresponding author, E-mail: qiumin@westlake.edu.cn

Optical neural networks (ONNs) have emerged as a groundbreaking paradigm in the pursuit of next-generation computing architectures, offering unprecedented advantages in terms of speed, energy efficiency, and parallelism. Rooted in the unique physical implementation principle of "architecture-as-function, propagation-as-computation", ONNs leverage the intrinsic properties of light—such as large bandwidth, low latency, and minimal energy consumption—to perform computational tasks in a manner fundamentally different from traditional electronic systems. This novel computing paradigm has attracted widespread attention from both academia and industry, as it presents a promising solution to the growing computational demands imposed by large-scale artificial intelligence (AI), especially in the context of artificial general intelligence (AGI) and real-time decision-making systems. However, despite their theoretical and experimental progress over the past decade, practical ONNs still face substantial technical barriers, particularly in scaling up both hardware and software components for effective training and inference.

One of the most significant challenges in the development of ONNs has been the reliance on external, GPU-based systems for training procedures, which undermines the inherent speed and energy efficiency of optical computation. Addressing this critical limitation, a research team led by Professors Lu Fang and Qionghai Dai at Tsinghua University has made a pioneering breakthrough by introducing a fully forward-mode training methodology for ONNs. This innovative approach enables *in-situ* training within the optical domain, eliminating the need for off-chip computational resources and paving the way for truly autonomous optical learning systems. Their work culminated in the design and implementation of the "Taichi" chiplet—a large-scale, general-purpose intelligent optical computing chiplet that integrates a distributed, broad-range optical computing architecture. The Taichi chiplet demonstrates remarkable capabilities in handling complex neural network operations, including both training and inference, thereby significantly advancing the scalability and applicability of photonic computing systems.

Recognized as one of China's Top 10 Scientific Advances in 2024, the Taichi project not only marks a pivotal milestone in optical computing but also highlights the transformative potential of photonic systems in reshaping the future of intelligent computation. This paper offers a comprehensive overview of the evolution of optical neural networks, tracing their development from early theoretical models to recent experimental implementations. It delves into key advancements in ONN architecture design, training methodologies, and diverse application domains, with a particular emphasis on the technical innovations embodied in the Taichi chip. Furthermore, the paper critically examines the current state of the field, identifies remaining challenges such as material limitations, noise sensitivity, and system integration, and outlines promising research directions.

This paper explores the future trajectory of optical computing, particularly in the context of industrial adoption and integration with existing AI ecosystems. It discusses the potential convergence of ONNs with quantum computing, neuromorphic engineering, and hybrid photonic-electronic systems, offering insights into how optical computing may redefine the architecture of intelligent systems in the post-Moore's Law era. By synthesizing theoretical foundations, recent breakthroughs, and forward-looking perspectives, this review aims to serve as a valuable reference for researchers and practitioners seeking to understand and contribute to the rapidly evolving field of optical neural networks.

Optical computing, optical neural network (ONN), optical computing architecture, large-scale intelligent photonic computing, optoelectronic intelligent computing, artificial intelligence (AI)