基于时域计算的高吞吐量光学神经网络

Shuang Zheng[1,2,3,4,†], Jiawei Zhang[1,2,3,4,†], Weizhen Yu[1,2,3,4] & Weifeng Zhang[1,2,3,4,*]

[1]北京理工大学信息与电子学院雷达研究实验室，中国北京 100081

[2]教育部卫星导航电子与信息技术重点实验室（北京理工大学），中国北京 100081

[3]北京理工大学重庆创新中心，中国重庆 401120

[4]重庆市新型民用雷达重点实验室，中国重庆 401120

[†]这些作者贡献相同：Shuang Zheng、Jiawei Zhang。

*通讯及材料请求请联系 Weifeng Zhang。

（电子邮箱：weifeng.zhang@bit.edu.cn）

摘要

一种新兴的、基于神经网络的生成式人工智能（AI）正日益普及，凭借变革性的能力创造新颖且

并且可按需设定加权系数，从而使光神经网络具有很强的可编程性。通过利用波长复用技术，可

引言

神经网络是一种计算模型，由分层结构中相互连接的节点或神经元构成，用于处理信息，其工作

（AI）任借创造全新原创内容的变革性能力，开始广受欢迎[2-6]。为了处理更复杂的模式，人们开发、训练并维护包含数百万至数十亿参数的大型生成式模型。因此，计算基础设施亟需具备海量且高能效的计算能力[7, 8]。用于摩尔定律难以持续，且冯·诺依曼架构存在局限，传统数字计算机正面临难以跟上的困境[9, 10]。模拟光计算利用光子的独特属性，包括宽带宽、低延迟和高能效，被视为一种有望光当计算加速器的解决方案[11, 12]。

近年来，人们投入大量精力研究光学神经网络（ONNs），并提出了多种实现方案[13, 14]。

自由空间光学或集成光学的光学加权或和实验实现方案见诸报道。在自由空间光学中，输入平面和输出平面上的空间光强分布可直接视为输入向量和输出向量。沿光传播方向放置多层级联的衍射光学元件（DOEs）或空间光调制器（SLMs），并在其中编码权重系数向量。光经过DOEs或SLMs后，可用透镜或透镜阵列收集平面上的光强，作为加权求和的输出[15-18]。得益于DOEs和SLMs的像素可重构性，输入光强图案和加权系数能够灵活调节，这对于提升所得光学神经网络的

与自由空间光学相比，集成光学在大规模集成、超小占用空间和大幅降低能耗方面具有巨大

中得到了广泛研究。基于硅的集成光学神经网络方案也在近期被频繁提出。

的研究中 [19-30]。典型的片上光学神经网络配置采用级联马赫–曾德尔干涉仪（MZI）构成的网[illegible]

[illegible]

尽管已有多种基于自由空间光学或集成光学的神经网络得到展示，但其主要问题在于，光神[illegible]

[illegible]

本文提出并展示了一种基于时域计算的高吞吐量 光学神经网络。核心加权求和运算通过高速[illegible]

得益于时域计算的高速特性，实验展示了光神经网络的高数据吞吐量；此外，加权系数可以按需

能力和高度可重构性的可拓展光神经网络，这为未来的大规模人工智能应用展现了巨大潜力

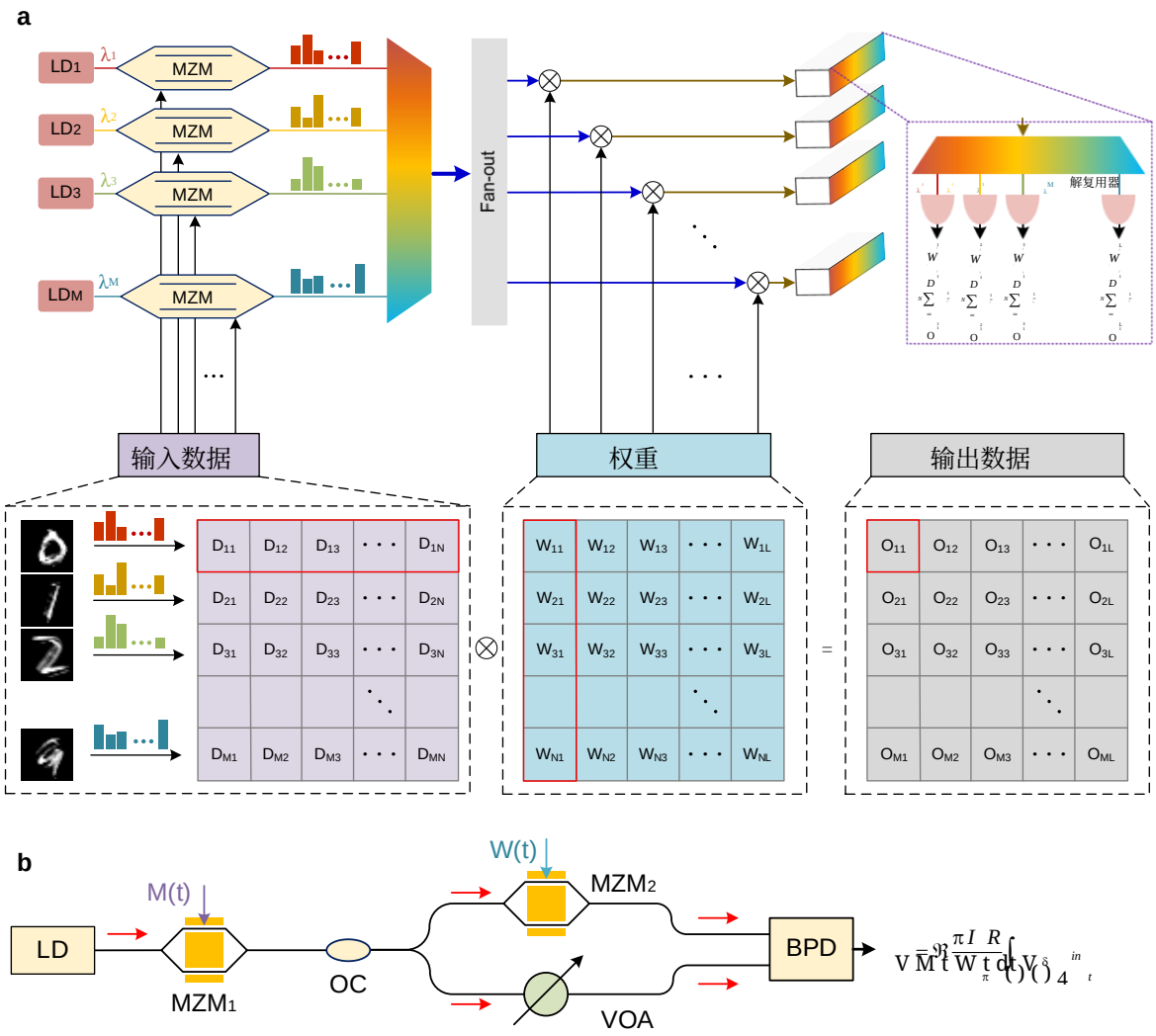

图1 （a）所提出的基于波长和空间复用的高吞吐量时域光计算系统示意图。LD，激光二极管；

结果

得益于光学提供的复用技术，图 1(a) 提出了一种基于波长和空间复用的高吞吐量时域光计算系统，[illegible]

[illegible]

图1(b)给出了所提出的时间光子加权求和运算单元的原理。输入图像数据 被展平为实值向量，[illegible]

[illegible]

数据信号预先调整为 。

$$M_L(t)=\frac{2V_\pi}{\pi}\arcsin\left[\sqrt{M(t)}\right] \tag{1}$$

其中 是马赫-曾德尔调制器$_1$的半波电压。

因此，在马赫-曾德尔调制器的输出端$_1$，，光信号强度可表示为：

$$I_{MZM1}(t)=I_{in}\sin^2\left[\frac{\pi}{2V_\pi}M_L(t)\right]=I_{in}M(t) \tag{2}$$

其中 为输入光信号强度。

随后，调制光信号通过一个 3 dB 光耦合器分为两路。在上路中，光信号被导入另一个高速电光强度调制器 MZM$_2$ 其中，展平为向量的加权系数被加载到光信号上。通过精确控制

两个马赫-曾德尔调制器中数据和权重加载的同步时间，即可完成数据向量与权重向量之间的乘法运算。由于马赫-曾德尔调制器$_2$偏置在正交传输点，马赫-曾德尔调制器$_2$输出端的光信号强度可写为:

$$I_{upper}(t)=\frac{I_{MZM1}(t)}{4}\left[1+\frac{\pi}{V_{\pi}}W(t)\right] \tag{3}$$

在下方路径中，光信号经过一个VOA，其强度可以调节。

$$I_{lower}(t)=\frac{\alpha}{2}I_{MZM1}(t) \tag{4}$$

其中 是 VOA的衰减系数。 随后，使用平衡光电探测器（BPD）收集来自两条路径的光信号，并

$$i_{\mathrm{BPD}}(t)=\Re\frac{I_{in}M(t)}{2}\left[\frac{1}{2}+\frac{\pi}{2V_{\pi}}W(t)-\alpha\right] \tag{5}$$

其中 是BPD的响应度 。当衰减系数 等于1/2时，该方程可简化为: α

$$i_{\mathrm{BPD}}(t)=\Re\frac{\pi I_{in}}{4V_{\pi}}M(t)W(t) \tag{6}$$

由于BPD具有低速特性，商用BPD中存在固有的RC电路，可用于执行时间累加。因此，在

$$V=\Re\frac{\pi I_{in}R}{4V_{\pi}}\int_{\delta t}M(t)W(t)dt \tag{7}$$

其中 是 寄生电阻， 是时间累加窗口。

由公式7可知，采用所提出的方法可以实现加权求和运算。将输入数据矩阵和权重矩阵展平

调制带宽的限制。据报道，目前市场上可用的调制器通常具有高达50 GHz的调制带宽[24]。从而能够实现高速乘法。此外，调制器的高速特性也带来了权重可重构性强这一附带优势。为进行求和，我们采用低速平衡光电探测器（BPD）。这带来一重优势。首先，得益于其差分输出，通过组合两个光电探测器（PD）输出中的共模部分，可以将光学求和结果提取为差分光电流，其次，

加权系数得以实现；第三，BPD中的电路为求和所需的时间累加提供了一种直接的解决方案。其

越长

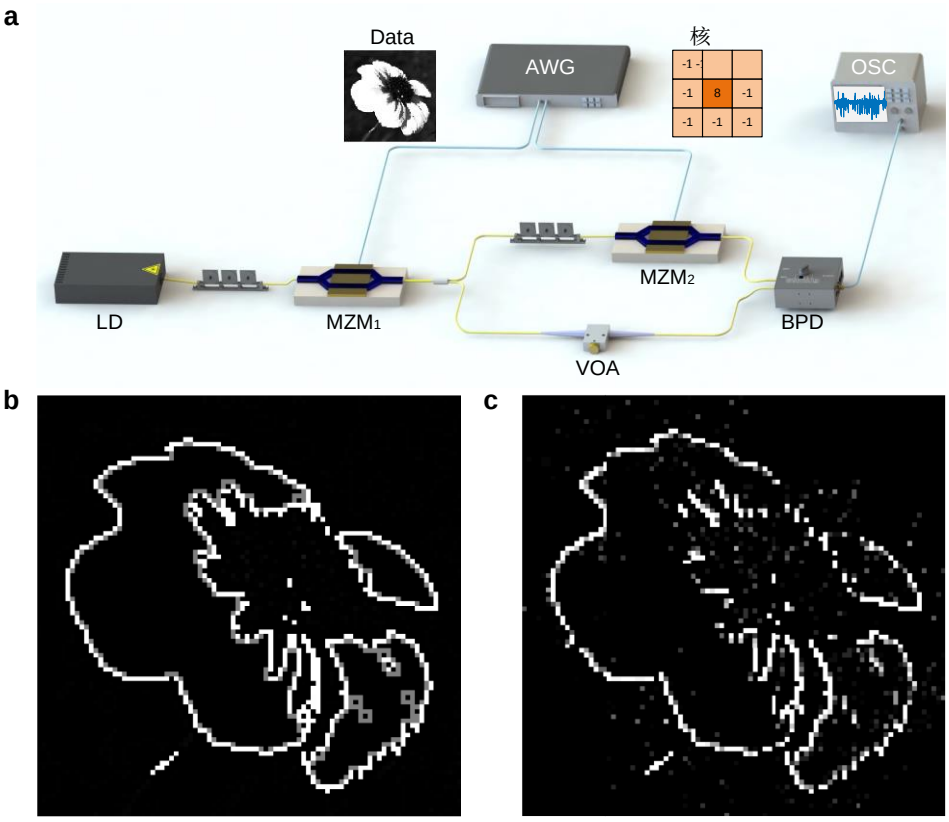

图2 (a) 用于图像边缘检测处理的光子加权求和运算单元的实验设置。LD，激光二极管；MZM，马赫-曾德尔调制器；VOA，可变光衰减器；BPD，平衡光电探测器；AWG，任意波形发生器；OSC，示波器。(b) 边缘检测图像的理论结果。(c) 边缘检测图像的实验结果。

图2(a)展示了所提出的时间光子加权求和运算单元的实验设置。使用 92×92 花朵图像作为输入数据，并采用 3×3 拉普拉斯核作为图像边缘检测任务所指定的加权系数。使用双通道任意波形发生器（AWG）分别按时间序列生成数据向量和核向量。将两个高速马赫-曾德尔调制器级联，以执行乘法的向量加载。带宽为150 MHz的平衡光电探测器（BPD）实现光电探测和时间累加。使用实时示波器记录输出时间电信号，并恢复边缘检测图像。图2(b)展示了使用数字信号处理算法得到的边缘检测图像理论结果。图2(c)给出了使用光子时域计算得到的边缘检测图像实验结果。可以看出，

结果彼此吻合良好，这验证了所提出的 光子加权求和运算单元的有效性。

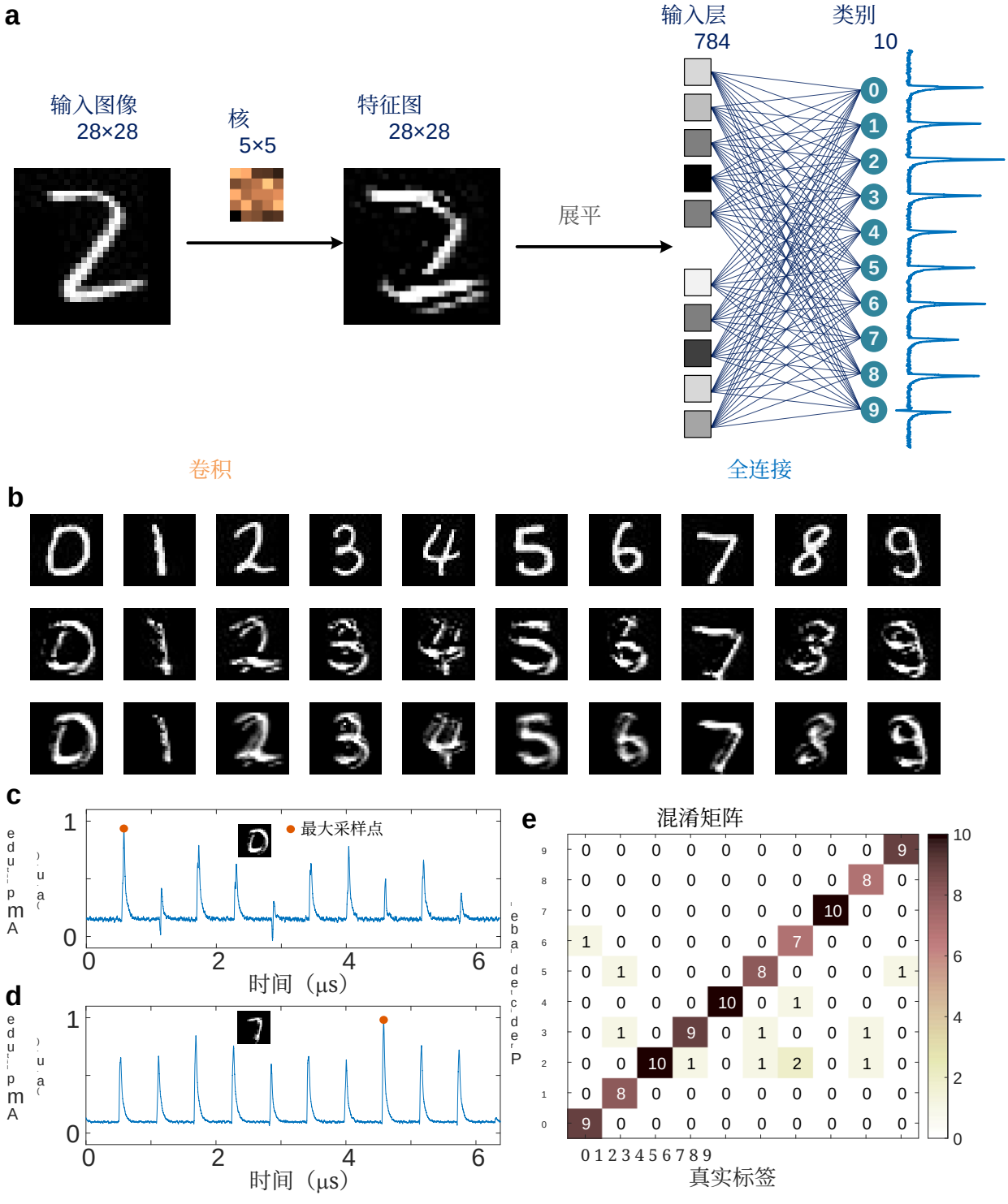

图 3 用于图像分类的 OCNN。a 基于 OCNN 的图像分类示意图。该

OCNN 由卷积层和全连接层组成。这两层中的线性运算均通过光子加权求和运算单元实现。b 理论计算结果与 10 个特征图的实验结果。c、d 全连接层测得的特征图及其对应的时间波形。e 通过 100 次重复实验获得的混淆矩阵。

随后，利用所提出的 光子加权求和运算单元构建光学卷积神经网络。卷积神经网络（CNN）作为一类人工神经网络，凭借强大的特征提取能力，在视觉图像分析中得到广泛应用 [35]。图 3(a) 展示了用于对 MNIST 数据集中的手写数字图像进行分类的典型卷积神经网络结构 。第一层为卷积层，用于执行特征学习；第二层为充当分类器的全连接层。由于在卷积层和全连接层中，均仅通过权重矩阵对输入向量执行线性变换，因此这两层可以利用所提出的光子加权求和运算单元实现。

图 3(b) 给出了基于所提出的光子加权求和运算单元的卷积层结果。第一行是待处理的原始手写数字图像，其中每一幅都是一张 28×28 手写数字图像。通过对输入图像执行线性乘加运算，利用一个 5×5 卷积核矩阵提取特征。运算结果被重构为特征图。第二行是 理论重构的特征图，第三行是实验重构的特征图。可以看出，光学计算得到的特征图与理论结果吻合良好，这再次验证了所提出的运算单元实现卷积层的有效性。

通过展平操作，每个重构的特征图都被转换为大小为 784 的向量。为了识别 0 到 9 的手写数字，将矩阵大小为 10×784 的全连接层训练好的权重库划分为 10 个长度为 784 的向量，并将其与特征图向量一同依次送入光子加权求和运算单元。全连接层的运算结果被记录为时间波形。图 3(c) 展示了加载数字 0 的理论特征图时全连接层的实验结果。波形中从左到右的 10 个峰值分别对应 0 到 9 的手写数字，其中数值最大的峰值表示所识别的数字。如图 3(c) 所示，最大采样点对应于手写数字 0。图 3(d) 展示了加载数字 7 的实验特征图时全连接层的实验结果，与理论计算结果一致。

我们进一步进行了更多手写数字分类的数值测试。图 3(e) 中 100 幅测试图像的混淆矩阵显示，生成的预测结果准确率为 88%，而在电子数字计算机上计算的数值结果准确率为 93%。这是由于系统

受到环境扰动和马赫-曾德尔调制器偏置漂移所导致的不稳定性影响，实验精度低于使用相同卷积神经网络模型理论计算得到的精度。通过采用片上集成和反馈控制，时域计算系统将更加稳定，准确率也有望进一步提高。得益于时域计算的高速特性，实验验证了光神经网络具有较高的数据吞吐量，并且可以按需指定加权系数，从而使光神经网络具有很强的可编程性。

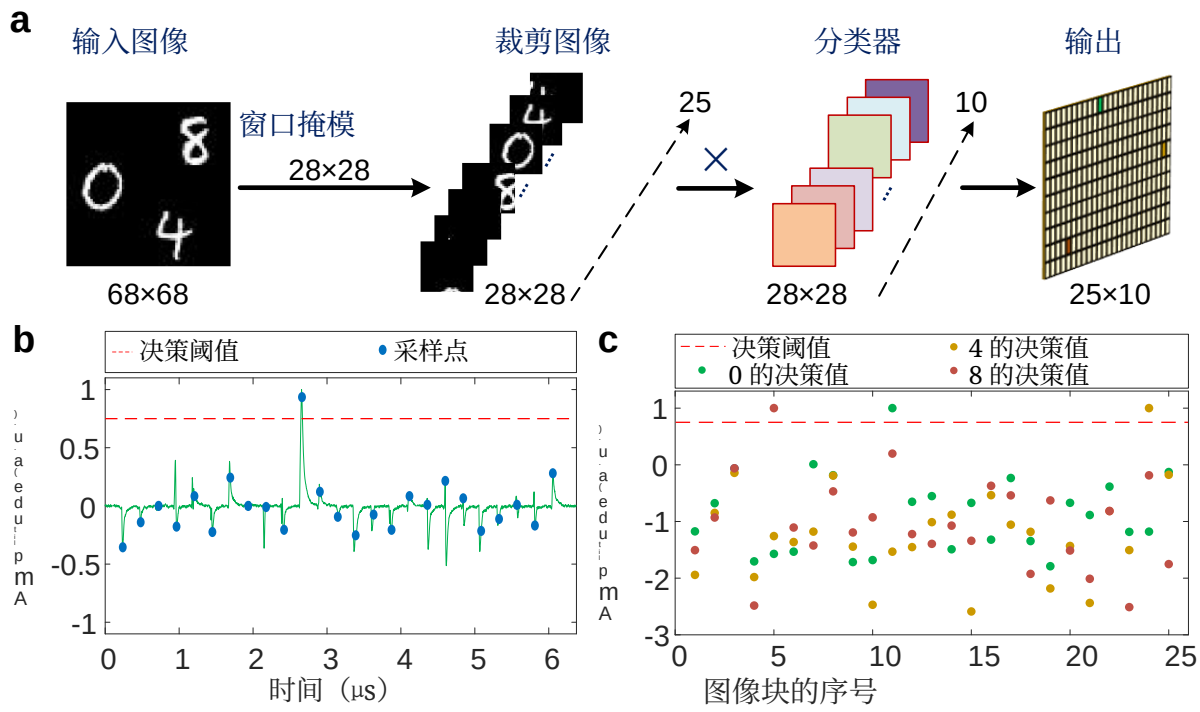

图5 滑动窗口目标检测。a 在 68×68 图像中检测手写数字的示意图。b 检测手写数字 0 时，在平衡光电探测器输出端记录的时间波形。c 对 25 个裁剪图像与三个分类器矩阵之间的加权求和结果进行汇总，用于检测数字 0、4 和 8。

采用光子加权求和运算单元进行滑动窗口目标检测。滑动窗口方法是目标检测中一种非常经典的方法，为其他先进的目标检测算法奠定了坚实基础 [37, 38]。图 5(a) 展示了在图像中使用滑动窗口检测多个手写数字的示意图。固定的 28×28 窗口系统地滑过整幅图像后，尺寸为 68×68 的图像按时间序列被裁剪为 25 个图像块。每个裁剪图像依次输入线性分类器，该分类器由 10 个尺寸为 28×28 的矩阵组成。每个矩阵都经过充分训练，用于分类特定数字。每个裁剪图像与每个矩阵之间均执行加权求和运算。在分类器的输出端，会生成一个尺寸为 25×10 的矩阵，其中每个元素都是每个裁剪图像与分类器中每个矩阵之间的加权求和结果。输出矩阵中的元素表示一个决策值。通过选择

超过决策阈值的决策值，可以检测出相应的手写数字，并根据输出矩阵在输入图像中定位其位置。

利用所提出的光子加权求和运算单元，可执行滑动窗口多数字检测。图 5(b) 显示了将 25 张裁剪图像展平并输入 MZM1，同时将用于检测手写数字 0 的第一个分类器矩阵加载到 MZM_2 时，平衡光电探测器输出端记录的时间波形。绿色曲线表示 25 张裁剪图像与第一个分类器矩阵之间的加权求和结果。虚线红线表示的决策阈值用于判断图像块中是否存在数字 0。超过决策阈值的采样点表明了物体在原始图像中的位置。可以看出，只有一个决策值大于阈值，这证实了手写数字 0 的存在。通过确定采样点位置，可以在11^{st} 裁剪图像中定位该数字，而该图像对应于原始输入图像中的特定位置。

图 5(c) 汇总了 25 张裁剪图像与分别用于检测数字 0、4 和 8 的三个分类器矩阵之间的加权求和结果。可以看出，当决策值大于阈值时，即可检测到手写数字。通过确定对应的裁剪图像编号，可以定位数字在原始输入图像中的位置。得益于所提出的基于时域计算的光子加权求和方法具有高吞吐量，可执行滑动窗口多数字检测。

兼容WDM的并行光子计算。 通过利用波长复用技术（WDM），可以构建具有强大计算能力和

为概念验证实验演示，研究人员构建了一种兼容WDM的全连接神经网络，采用两个波长复用通道，实现MNIST 数据集中的手写数字图像分类。图6展示了基于兼容WDM的并行光子计算的并行全连接神经网络示意图。由两幅28×28 图像展平得到的两个1×784 向量通过两个与狮-曾德尔调制器（MZM3）加载到两个光载波上。随后，两个调制光信号通过光耦合器合并，并发送至分

光结果。

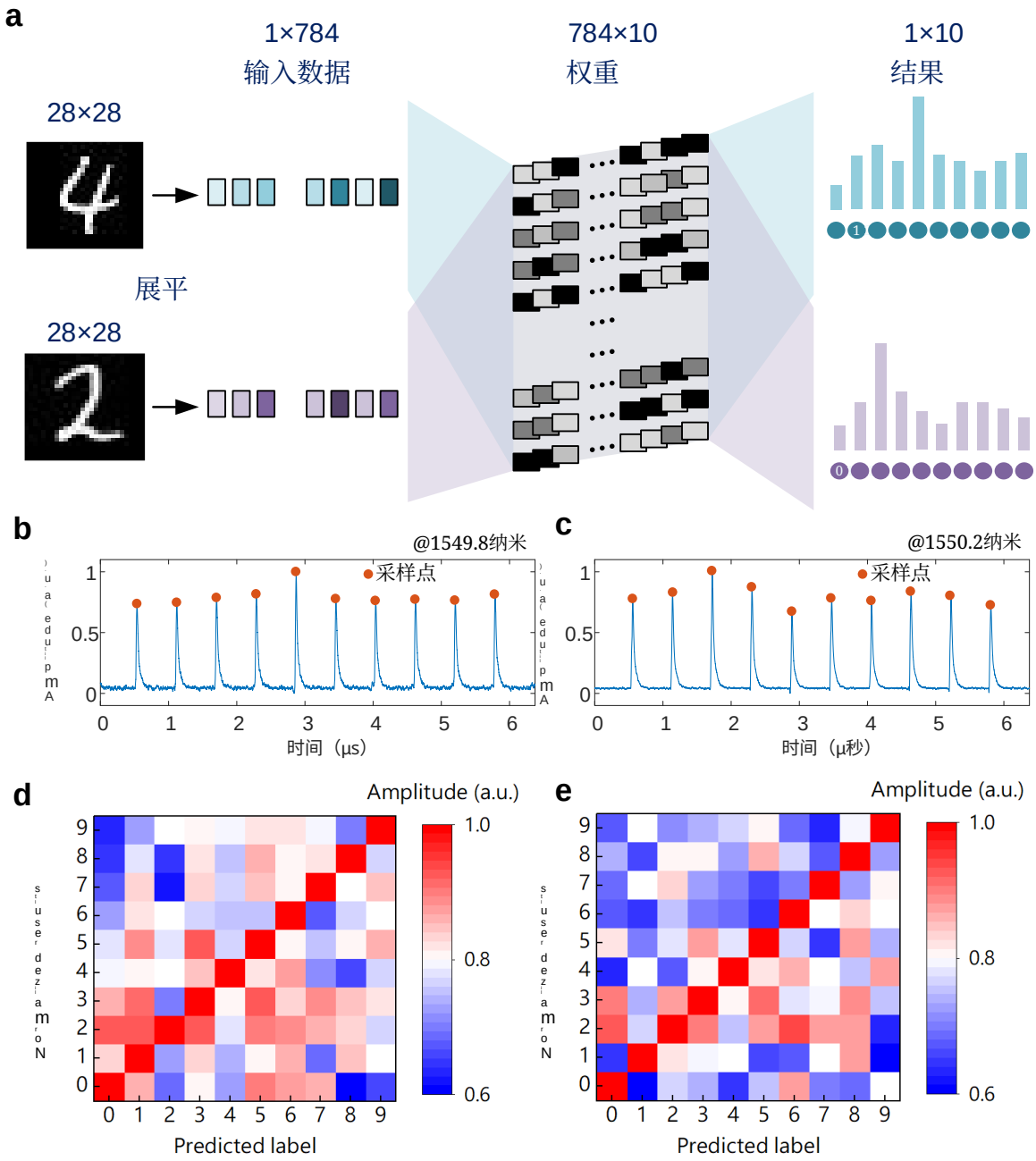

图6 用于 图像识别任务的并行全连接神经网络。a 基于兼容WDM的并行光子计算的可扩展全连接

1 1549.8纳米处的归一化采样结果。e 1550.2纳米处的归一化采样结果。

图6(b)和(c)分别展示了在1549.8纳米和1550.2纳米处，平衡光电探测器（BPDs）输出端测

表示权重求和结果的波形。所测时间波形中的最大采样点位置表征手写数字。将0至9的不同手写

波长于0至9手写数字的归一化采样结果，与理论计算结果高度一致，很差于光子时域计算，可引入波分复用技术以大规模增加通道数，从而显著提升光子计算系统的计算能力

表1 已报道的光学神经网络与本工作的比较。				
Year	**Ref.**	求和	架构	输入向量 **size**
2017	20	空间干涉	MZI阵列	4×1
2021	26	空间干涉	MZI阵列	4×1
2023	29	空间干涉	MZI阵列	4×1
2022	42	空间干涉	级联MMI	10×1
2021	23	多波长	光学频率梳	16×1
2021	41	多波长	光学频率梳	90×1
	Ours	时间积分	BPD	1×784

讨论

近年来，人们提出了多种不同方法，并通过实验演示了如何实现神经网络的光子加权求和运算。表1从求和实现方式、系统架构以及所实现神经网络的输入向量大小等方面，总结了此前已报道的工作和我们的工作。光学空间干涉是实现光功率求和的一种直接方法。利用MZI结构会产生光学空间干涉，通过调节两条臂之间的相位差，即可方便地实现光子加权求和。然而，由于其采用干涉结构，稳定性是一个关键问题。与分立式方案相比，更受青睐的方法是采用片上集成方案，通过级联多个MZI来实现多次空间干涉。其主要缺点是，MZI结构通常尺寸较大，达到数百微米，因此同一芯片上可集成的MZI数量有限。因此，要扩展基于MZI的光神经网络极具挑战性。从表中可以看出

迄今为止，片上 MZI 阵列一直用于实现较小的向量维度。多波长复用是实现光功率求和的另一种直接方法。每个波长对应一个输入元素，而波长数量决定了输入向量大小。得益于光学频率梳的快速发展，多波长光源已成为可能。关键问题在于波长数量仍然有限，这不可避免地导致输入向量大小受限，进而影响计算能力。

我们提出的方法利用时间积分实现光学求和。与人脑依据事件发生时序处理信息类似，时域计算将数据和权重编码到时间序列中进行乘法运算，并使用积分器完成求和。在实验演示中，我们实现了一个输入向量为 1×784 的全连接神经网络，展现出巨大的计算能力潜力。其主要优势包括高吞吐量的光子加权求和以及良好的可扩展性。数据吞吐量取决于调制器的带宽和 BPD 的时间累加窗口。采用最先进的光子发射器和光电探测器后，加权求和运算单元的速度可达 100 GOPS[43]。结合多波长复用，可扩展至 100 个并行波长计算通道，总有效计算速度为 10 TOPS。此外，值得一提的是，我们提出的时域计算架构的核心可集成在单个硅光子芯片上，从而降低功耗并显著增强可扩展性。

总之，我们提出了一种基于时域计算的高吞吐量光子计算架构。 其核心加权求和运算通过高速电光调制和低速平衡光电探测来实现。输入数据和权重分别编码为时间序列，并依次通过两个电光调制器加载到光信号上。通过精确控制数据和权重加载的同步时间，即可完成乘法运算。随后，信号经过平衡光电探测器，借助低速光电探测器内在电子积分电路中的电子累积完成求和。由此，便在光学域中基于时域计算实现了线性加权求和运算。利用所提出的光学线性加权求和，我们实现了全连接神经网络和卷积神经网络。得益于时域计算的高速特性，我们通过实验展示了光神经网络的高数据吞吐量；此外，加权系数可以按需指定，使光神经网络具有很强的可编程性。通过采用波长复用技术，我们构建了具备强大计算能力和高度可重构性的可扩展光神经网络，在未来的大规模人工智能应用中具有巨大潜力。

方法

实验设置。 实验中使用的激光源（NKT Koheras BASIK）发出波长为 1550 nm 的光信号。图像向量和权重矩阵由任意波形发生器（Keysight M8196A AWG）生成。电光调制通过富士通 FTM7937 调制器实现。采用增益箱带宽与切换的平衡光电探测器（Thorlabs PDB 450C）完成光电转换。实时示波器（Tektronix DPO75902SX）记录平衡光电探测器输出端的时间波形

参考文献

1. Krogh, A. 什么是人工神经网络? 《自然·生物技术》26, 195–197 (2008). 2. LeCun, Y., Bengio, Y. & Hinton, G. 深度学习。《自然》521(7553), 436-444 (2015). 3. Long, J.

10. Tallents, G., Wagenaars, E., & Pert, G. 极紫外波长光刻。《自然·光子学》4(12), 809- 811 (2010). 11. Caulfield, H. J. & Dolev, S. 为什么未来的超级计算需要光学。《自然·光子学》4(5)

23. Feldmann, J., 等. 使用集成光子张量核心进行并行卷积处理。《自然》589(7840), 52-58 (2021). 24. Huang, C. 等. 用于光纤非线性补偿的硅基光子电子神经网络。《自然·电子》4(11)

35. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. 卷积神经网络综述：分析、应用与展望。IEEE Trans. Neural Netw. Learn. Sys. (2021). 36. Xu, X. 等. 基于克尔微梳的高速、可扩展光学神经

补充信息

补充信息链接在论文的在线版本中。

作者贡献

S.Z.、J.Z. 和 W.Z. 提出了这一想法。J.Z. 和 S.Z. 完成了实验。S.Z. 和 J.Z. 分析了数据。S.Z. 和 J.Z. 参与了实验结果的讨论。W.Z. 负责监督并协调所有工作。所有作者共同撰写了稿件。

致谢

本研究得到国家重点研发计划（2019YFB2203303）和国家自然科学基金委员会（国家自然科学基金委员会）（62105028和62071042）的支持。

数据可用性

支持本研究结果的数据可在合理请求后向通讯作者获取。

竞争性利益

作者声明不存在竞争性利益。