光学编码器的光子学优势

作者：Luocheng Huang[1], Quentin A. A. Tanguy[1], Johannes E. Fröch[1,2], Saswata Mukherjee[1],Karl F. Böhringer[1,3,4], Arka Majumdar[1,2,*]

所属机构：

[1]华盛顿大学电气与计算机工程系，西雅图，华盛顿州 98195

[2]华盛顿大学物理系，西雅图，华盛顿州 98195

[3]华盛顿大学生物工程系，美国华盛顿州西雅图，98195

[4]纳米工程系统研究所，华盛顿大学，美国华盛顿州西雅图，98195

* 通讯作者。电子邮箱：arka@uw.edu

摘要：

光学编码器的光子优势光能够并行执行大规模线性运算，这一特性近期促成了众多光学辅助人工神经网络（人工神经网络）的演示。然而，在系统层面上，光学相较于纯数字人工神经网络的明确优势尚未得到证实。尽管光学确实能够非常高效地执行线性运算，但由于缺乏非线性和信号再生，光学与电子学之间的信号转换需要高功耗、低延迟。此外，激光器和光电探测器也需要消耗大量功率，而计算能耗时往往会忽略它们。在这里，我们没有将传统数字运算映射到光学领域，而是共同优化了一种工作于非相干光下的光学-数字混合人工神经网络，因此它适用于在环境光下运行。在纯数字人工神经网络与混合光电人工神经网络的延迟和功耗保持一致的情况下，我们确定了一个低功耗/低延迟范围，在此范围内，光学编码器的分类准确率高于纯数字人工神经网络。然而，在该范围内，总体分类准确率低于采用更高功耗和延迟时所能达到的水平。我们的结果表明，在可放宽人工神经网络总体性能、优先考虑低功耗和低延迟的应用中，光学技术可能优于数字人工神经网络。

一句话总结：我们清晰地展示了，在低延迟和低功耗约束下，协同设计的光学编码器与数字后端的性能优于纯数字人工神经网络。

正文：

简介：

在过去十年中，人工智能（AI）和深度学习领域取得了加速发展，揭示了人工神经网络（人工神经网络）在各种应用中的潜力和能力。近期的演示甚至以聊天软件和艺术渲染程序的形式进入公众视野。它们近期的成功可追溯到计算算法和图形处理器（图形处理器）等数字硬件方面的重大突破（1）。尽管成果令人瞩目，但随着人工神经网络规模的扩大，深度学习数字实现的功耗和延迟扩展表现出不利影响。这对人工神经网络的进一步扩展（2、3）以及其在低功耗、实时问题中的适用性构成了严重限制。

光可能是解决这一扩展问题的答案。得益于其固有的并行性、高速度和模拟特性，光为构建高能效、高速度的人工神经网络（ANN）提供了一种极具吸引力的替代方案，以取代电子实现。早在这一领域发展的初期，人们就已经认识到了这一点，并且早在20世纪90年代，就有多项实验报告了光学人工神经网络（4, 5）。遗憾的是，由于技术和基础理论方面的原因，相关研究进展陷入停滞，这些原因大体上可分为内在问题和外在问题。光学领域的内在问题包括：光学元件体积庞大且对失准的容忍度较低；空间光调制器的空间带宽积有限；以及缺乏非线性激活机制。外在问题则源于人们对人工智能算法和自适应学习的理解不足，以及电子计算系统的迅猛崛起。

鉴于当前电子硬件的局限性以及我们对人工智能认识的加深，外在问题在一定程度上得到了缓解。与此同时，纳米制造设施的发展以及先进电磁仿真器的普及，推动了多功能纳米光学器件的大规模制造，例如平面超光学（6, 7）和集成光子器件（8）。新兴材料体系与这些纳米光学结构相结合，使得制造出类似于电子集成电路的单片光子集成电路（PIC）成为可能（9）。纳米光子学和人工智能领域的这些创新，加上数字方式实现人工神经网络的严重局限，近年来引发了人们对重现光学辅助人工神经网络的浓厚兴趣（10-17）。

然而，迄今为止，尚无已发表的研究明确证明，在推理方面，光学相较于数字人工神经网络具有优势。大多数实现仅展示了用光学对应方案替代一小部分线性部分，其余部分仍由数字电子学处理。尽管光学在实现小型子系统（通常是线性部分）方面具有明显优势，但完整人工神经网络的功耗和延迟还包括光学域与电子域之间的信号转换（18），也就是探测器读出功率和激光功率，而这些因素往往被忽略。事实上，考虑这些能耗的分析表明，仅用光学实现一个卷积层并无优势，除非输入维度非常大（18）。此外，大量研究展示了针对极其简单的“玩具”问题进行分类，而这些问题没有数字基准（13, 14）。将专用光学人工神经网络的功耗和延迟与图形处理器（针对通用运算进行了优化）进行比较并不公平。降低数字人工神经网络功耗和延迟的方法有很多，例如用 XNOR运算替代矩阵乘法（19）。此外，还有许多剪枝算法可减少推理所需的运算次数。因此，迄今尚无明确证据表明，光学辅助人工神经网络相较于针对特定问题优化的纯数字框架具有优势。一个挑战在于，人工神经网络的计算复杂度无法被精确定义，因此数字部分功耗和延迟的精确计算取决于训练和技术。

在此，我们提出一种框架，用于精确比较纯数字人工神经网络与混合光电人工神经网络的推理性能。对于这两种人工神经网络，我们确保其功耗和延迟相同，因此通过比较分类准确率，可以清楚地评估它们各自的优势。图1展示了两种方案的示意图：纯数字方案和光学-数字混合方案。我们以非相干光对输入进行编码，因为人工神经网络的光学前端可利用环境光工作，无需额外能耗。在纯数字方案中，基于透镜的传感器在非相干光照射下采集物体图像，随后将图像传输至数字人工神经网络。在混合方案中，我们使用一种经过设计的光学器件——即光学编码器——取代透镜，以不同的基底采集图像，并将数据发送至数字后端。我们没有在光学系统中实现卷积运算等数字子系统，而是通过端到端设计框架，对光学前端（由亚波长衍射超光学实现）与数字后端进行协同优化（详见补充材料 S1、S3）(20)。

两种方案中，数字人工神经网络所采用的拓扑结构和资源（即节点、层和非线性函数的数量）保持不变，但权重和偏置有所不同。因此，我们确保两种方案的延迟和功耗保持相同。

在此，我们测试了 MNIST数据集在不同 N 值下的分类准确率。N 表示通过透镜或光学编码器在传感器中捕获的图像经分箱后的尺寸。随着发送至数字人工神经网络的数据输入维度 N 增大，延迟和功耗也随之增加。我们发现，两种情况下的分类准确率都会提高；但当 N 较大时，光学前端并无优势。不过，当 N 较小时，系统功耗和延迟也较低，此时采用混合光电人工神经网络可提高验证准确率（~10%）。我们通过实验验证了理论模型。我们的研究清楚地展示了人工神经网络推理中的光子学优势，尽管只有在整体系统性能低于可实现的最高性能时，才能观察到这种优势。

结果:

我们的数字后端由三个全连接层组成：$N \times 256$ （输入）、256×256 （隐藏）和 256×10 （输出）。前两层后面各接一个线性整流单元（ReLU）非线性函数，输出层则采用 Sigmoid 非线性函数。对于纯数字方案，我们通过对像素取平均值，将每张图像转换为 N 像素图像。我们选择了 8 种不同的 N ，范围从 1 到 100，以评估系统在数据输入量逐渐增加时的性能。我们通过反向传播损失函数来训练数字网络，该损失函数定义为输出与真实标注之间的交叉熵。在仿真中，我们获得了高达 ~98% 的验证分类准确率（详情见补充材料 S2）。值得注意的是，在先前的工作中，为了在 MNIST 数据集上达到相近的准确率，使用了多个层（17）；我们认为这是训练效率低下所致。对于混合式方案，我们使用亚波长衍射超光学对光学前端进行建模，不过也可以采用任意自由曲面光学表面来实现。具有不同输出维度的已制造光学前端如图2（b）所示。我们按照先前用于成像（20）的类似框架，使用相同的神经网络拓扑结构训练超构光学和数字后端（详见方法部分）。正如预期，随着 N 增加，分类准确率也随之提高。我们还发现，对于 $N > 8 \times 8$, 数字人工神经网络和混合人工神经网络的分类准确率相同。然而，当 N 较低时，混合人工神经网络的分类准确率超过数字人工神经网络。图3（a）展示了分类混淆矩阵示例，对比了实验验证

准确率：具有相同输入大小 $N = 3 \times 3$ 的混合人工神经网络与数字人工神经网络之间的准确率。从理论上看，加入光学前端后，分类准确率最高可提升 ~20%。图3（b）展示了验证准确率对比图。值得注意的是，即使只向数字后端输入一个数据点，我们预计采用光学前端仍能获得更高的分类准确率。这是因为对于 8位精度传感器，该单个输入可以有 256 种不同取值。

为验证这一设计，我们制备了超构光学器件（详情见补充材料S4），并通过实验测量其性能；实验中，我们使用OLED显示屏以绿光显示MNIST数据集中的图像（详情见补充材料S5）。非相干绿光穿过超构光学器件后，我们以8位精度在传感器上捕获数据。随后，我们对捕获的图像进行分箱处理，生成传递至数字后端的 N 个数据点。图2(c)中的一个实验样本展示了 3×3 编码器的信号处理过程。由于制造缺陷和错位，我们使用捕获的数据重新训练了数字后端（拓扑结构保持不变）。对于 $N \geq 3 \times 3$，我们的实验结果与理论预测非常吻合。需要指出的是，为 $N = 8 \times 8$ 优化的超构光学器件已损坏，因此我们无法采集相关数据。在较小的 N 下，实验结果与理论预测之间的偏差归因于实验噪声。尽管单个点能够为数字后端提供更多信息，但它会受到量化噪声的干扰，削弱光学编码器的作用。因此，我们得到的分类准确率与纯数字后端的预期结果相近。

讨论：

通过采用非相干光源和超构光学前端，我们构建了一个框架，使我们能够在同等条件下比较数字人工神经网络与光学辅助人工神经网络的性能。在两种情况下保持功耗和延迟不变时，我们展示了光学编码确实能向数字后端提供更多信息，使实验中的分类准确率提高了 ~10%。虽然我们的结果主要适用于 MNIST 数据集，但我们认为，这表明了在何种条件下光学前端有助于提升人工神经网络的性能（更多讨论见补充材料S6）。如果不对延迟和功耗作任何限制，就可以任意提高 N，并且总能找到优于混合式方案的数字解决方案。对此的一种解释是，任何光学实现都可以用数字方式建模，因此在没有任何限制的情况下，可以找到准确率与其光学对应方案处于同一数量级或更高的数字解决方案。较高的

多篇报告中光学辅助人工神经网络较高的分类准确率，很可能是全数字人工神经网络训练不足的体现。然而，在延迟或功耗受限的情况下，我们需要采用 N 的中间值；此时，光学前端能够提供更高效的解决方案，尽管其总体准确率较低。

致谢：

资金支持：

美国国家科学基金会（NSF-ECCS-2127235）；美国国防高级研究计划局（合同编号：W31P4Q-21-C-0043）

作者贡献：

概念构思：AM；方法学：LH、SM、QT、JF；研究：LH、SM、QT、JF；制备：QT；可视化：LH、QT；资金获取：AM、KB；项目管理：AM、KB；监督指导：AM、KB；撰写——初稿：LH、AM、JF；撰写

_ 审阅与编辑：LH、SM、QT、JF、KB、AM

利益冲突：AM 和 KB 是 Tunoptix 的共同创始人，该公司正在将类似技术商业化。

数据和材料可用性：所有数据均可在正文或补充材料中获取。

补充材料

材料与方法

补充文本

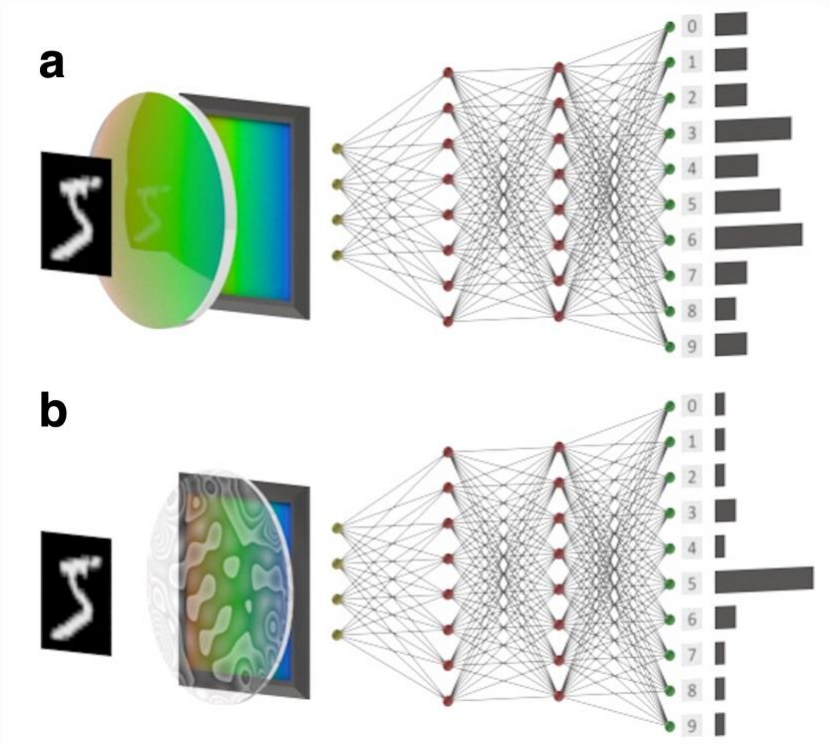


图1. 光学编码器与纯数字神经网络示意图。(a) 纯数字人工神经网络对带透镜传感器捕获的图像进行处理。(b) 无需使用透镜，经过设计的光学系统即可对捕获数据执行额外的线性运算。两种情况下，传感器的功耗和延迟都相同。使用资源相同的数字计算后端（层数和神经元数量），可确保功耗和延迟相同；二者均随数字后端输入数据的维度单调变化。

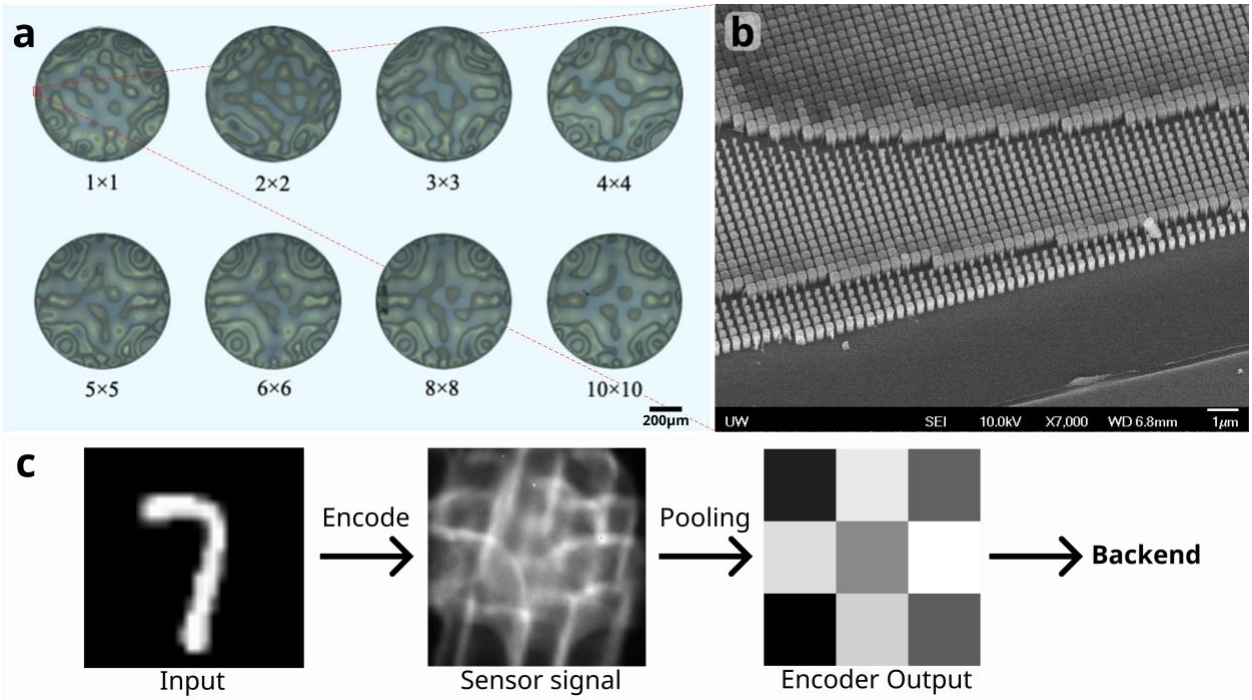


图2. 元光学编码器的制备与表征：(a) 不同输入尺寸下元光学编码器的光学显微镜图像。(b) 光学编码器的扫描电子显微镜（SEM）图像，所示区域为器件 1×1 上红框标示的区域。(c) 元光学编码器的实验输入、传感器信号和输出。

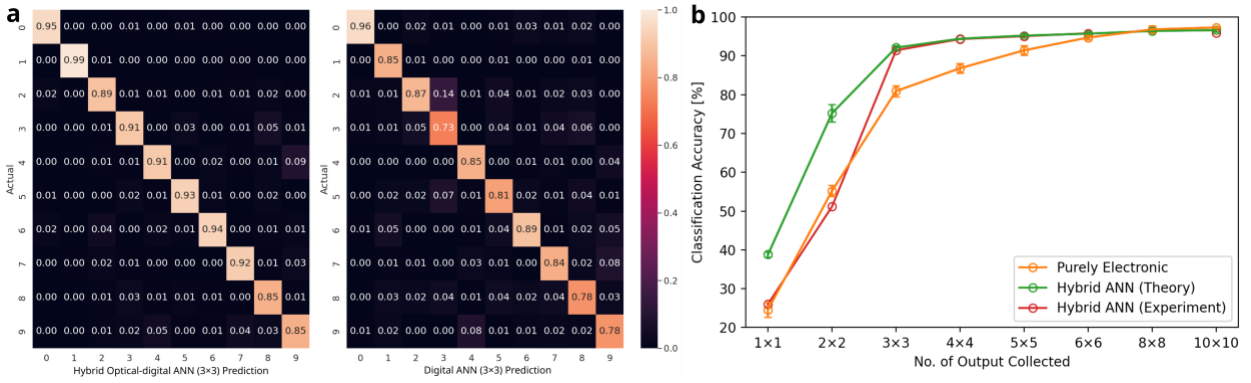


图3. 数字人工神经网络与混合人工神经网络的性能比较。(a) 在 $N = 3 \times 3$ 情况下，比较光学-数字混合人工神经网络与纯数字人工神经网络实验性能的混淆矩阵。(b) 纯电子

以及混合光电人工神经网络。误差棒表示一个标准差的范围。

参考文献：

1. Y. ABU-MOSTAFA，D. Psaltis，《光学神经计算机》。《科学美国人》256，88-95（1987）。2. N. H. Farhat，D. Psaltis，A. Prata，E. Paek，《霍普菲尔德模型的光学实现》。《应用光学》24，1469-1475（1985）。3. Y. LeCun，Y. Bengio，G. Hinton，《深度学习》。《自然》521，436（2015）。4. E. Strubell，A. Ganesh，A. McCallum，《现代深度学习研究的能源与政策考量》。第34届AAAI人工智能会议论文集，13693-13696（2020）。5. N. C. Thompson，K. Greenewald，K. Lee，G. F. Manso，《深度学习的计算极限》。arXiv:2007.05558，（2022）。6. A. Zhan 等，《低对比度介电超表面光学》。《ACS光子学》3，209-214（2016）。

7. N. Yu，F. Capasso，《采用设计型超表面的平面光学》。《自然·材料》13，139-150（2014）。8. L. Chrostowski，M. Hochberg，《硅光子学设计：从器件到系统》。（剑桥大学出版社，2015）。9. M. J. R. Heck 等，《混合硅光子集成电路技术》。《IEEE量子电子学精选主题期刊》19，6100117-6100117（2013）。10. Y. Shen 等，《利用相干纳米光子电路进行深度学习》。《自然·光子学》11，441（2017）。11. X. Xu 等，《用于光学神经网络的11 TOPS光子卷积加速器》。《自然》589，44-51（2021）。12. J. Feldmann 等，《使用集成光子张量核心进行并行卷积处理》。《自然》589，52-58（2021）。13. A. Sludds 等，《互联网边缘的分布式光子深度学习》。《科学》378，270-276（2022）。14. F. Ashtiani，A. J. Geers，F. Aflatouni，《用于图像分类的片上光子深度神经网络》。《自然》606，501-506（2022）。15. H. Zheng 等，《用于物体分类器的超光学加速器》。《科学进展》8，eabo6410（2022）。16. A. Ryou 等，《基于热原子非线性的自由空间光学神经网络》。《光子学研究》9，B128-B134（2021）。17. T. Wang 等，《每次乘法少于1个光子的光学神经网络》。《自然·通讯》13，123（2022）。18. S. Colburn，Y. Chu，E. Shilzerman，A. Majumdar，《用于卷积神经网络的光学前端》。《应用光学》58，3179-3186（2019）。19. M. Rastegari，V. Ordonez，J. Redmon，A. Farhadi，载于《计算机视觉 – ECCV 2016》，B. Leibe，J. Matas，N. Sebe，M. Welling 编。（施普林格国际出版，Cham，2016），第525-542页。20. E. Tseng 等，《用于高质量薄透镜成像的神经纳米光学》。《自然·通讯》12，6493（2021）。

补充材料：

光学编码器的光子学优势

作者：Luocheng Huang[1], Quentin A. A. Tanguy[1], Johannes E. Fröch[1,2], Saswata Mukherjee[1], Karl F. Böhringer[1,3,4], Arka Majumdar[1,2,*]

此 PDF 文件包括：S1. 元光学与数字后端的端到端设计 S2. 纯数字神经网络的设计 S3. 元光学的设计 S4. 元光学的制备 S5. 元光学编码器的实验细节 S6. 现有研究中光学优势的展示

S1. 超构光学与数字后端的端到端设计：混合光电人工神经网络

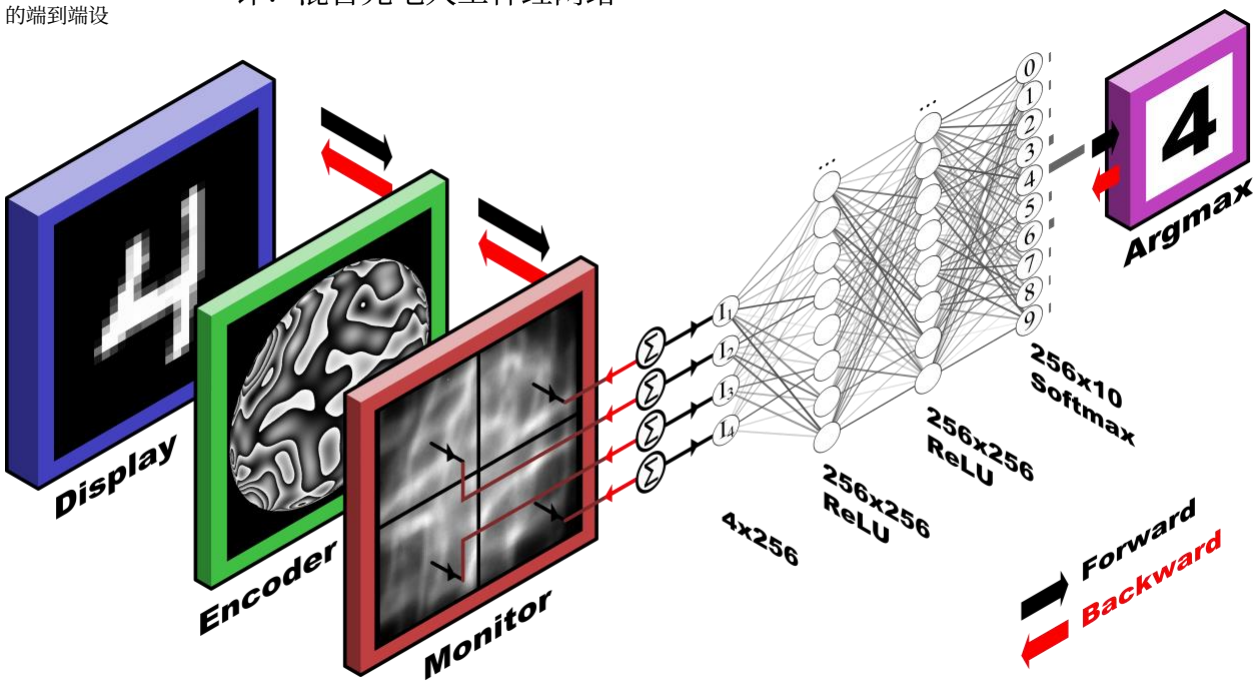


图 S1。混合光学-数字神经网络通过端到端可微流水线进行迭代设计。每次迭代包括损失的前向计算和损失的反向传播。此处展示了 2×2 编码的一个示例。

混合光学-数字神经网络的设计：

我们实现了一个端到端可微流水线，用于计算元光学编码器上的相位分布相对于混合光学-数字神经网络分类准确率的梯度。该流水线由三个阶段组成，分别是点扩散函数（PSF）仿真、成像仿真和分类。损失函数定义为混合神经网络输出与真实标注之间的交叉熵。我们使用 TensorFlow 2.8 作为自动微分引擎来实现前向计算。

在点扩散函数仿真阶段，法向入射的光通过超构光学传播，相位调制分布表示为 $\phi(x, y)$（图 S1 中的“编码器”）。相位分布由 z_j 参数化，即泽尼克多项式 R_n^m 的线性组合，其中：

$$R_n^m(\rho) = \sum_{k=0}^{\frac{n-m}{2}} (-1)^k \binom{n-k}{k} \binom{n-2k}{\frac{n-m}{2}-k} \rho^{n-2k}$$

我们使用 200 项泽尼克多项式对相位分布面进行参数化，使得 $\phi = \sum_{j=1}^{200} z_j R_j$，其中 $R_j \to R_n^m$，给定 $j = {}^{n}{}_{(n+2)_2}{}^{+m}$。

这些多项式经过预先计算后存入内存，并一直保留到整个优化循环结束。泽尼克多项式是正交基，可为空间超表面的相位调制提供空间正则化。我们发现，与优化每个空间位置的相位值相比，采用这类基函数可以避免优化过程陷入局部极小值。计算出相位分布后，我们使用带宽受限角谱法传播复场，以获得焦平面上的强度，即点扩散函数。

端到端流程的第二阶段是将点扩散函数与批量输入的 MNIST 图像进行卷积。首先，使用双线性插值对输入图像进行上采样，使其尺寸与点扩散函数强度数组相同。随后，这些输入图像会存储在内存中，供后续优化过程使用。接着，利用傅里叶卷积定理将点扩散函数与输入图像进行卷积，使得 $O = \mathrm{F}^{-1}\{\mathrm{F}\{PSF\} \cdot \mathrm{F}\{I\}\}$，其中 O 表示输出图像，I 表示输入图像。输出 O 可在图 S1 的“监视器”中查看。需要指出的是，这种成像建模方法仅适用于非相干发射条件下的成像，因为这正是我们旨在本文中展示的情形。

随后将输出图像输入可训练的数字后端人工神经网络，该网络包含多个顺序层，包括一个 $N \times N$ 平均池化层，以及两个各含 256 个单元的全连接层，

并采用 ReLU 激活函数；最后是一个含 10 个单元、采用 softmax 激活函数的全连接层，作为输出层。请注意，混合光电人工神经网络和纯数字人工神经网络采用相同的数字后端人工神经网络架构。请注意，$N \times N$ 等于数字后端的输入大小。数字后端的具体延迟和功耗取决于所用技术（软件、图形处理器或针对特定人工神经网络优化的 ASIC），但预计二者都会随着 N 的增大而单调增加。

前向计算后，我们得到人工神经网络输出与 MNIST数据集真实标注之间的交叉熵损失 L。随后，自动微分算法开始反向计算，在此过程中，我们得到损失相对于泽尼克多项式系数的梯度，即 $\partial \mathrm{L}/\partial z_j$。图 S1 中的红色箭头直观地表示了这一过程。然后，我们将该梯度应用于参数 z_j ，并乘以 Adam 优化器根据学习率 0.001 提供的因子。

训练循环包括损失函数的前向计算以及损失梯度的反向计算。该训练循环迭代执行，直至损失收敛。我们发现，200 次迭代足以实现收敛。前向计算和反向计算均在主频为 2.20 GHz 的 Intel Xeon 处理器上执行，并由配备 16GB RAM 的 Nvidia Tesla P100 加速。对于每种输入尺寸（$N \times N$），优化过程均运行了 4 次。

训练和验证准确率如图 S2 所示。训练混淆矩阵见图 S3。

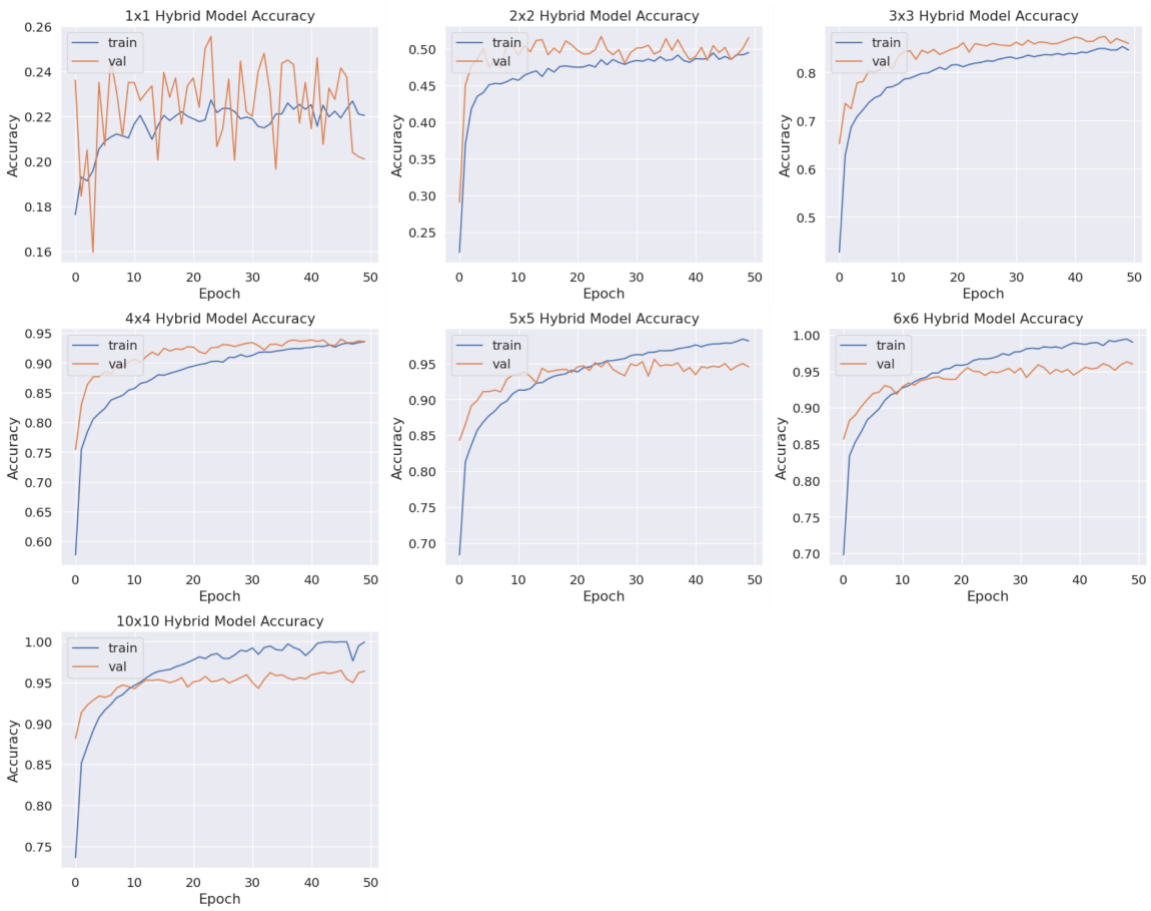


图 S2. 混合式人工神经网络分类的训练（train）和验证（val）准确率。

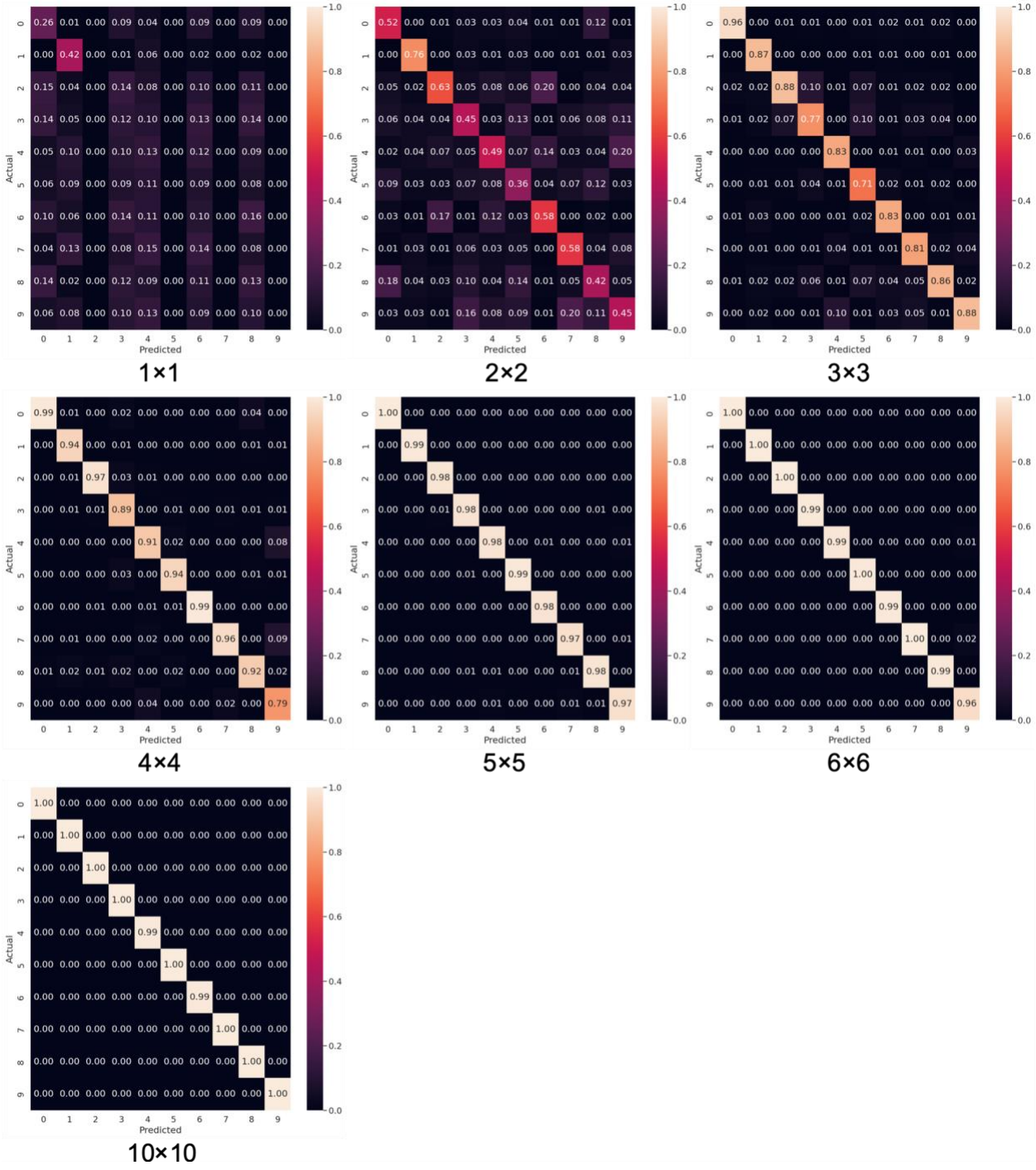


图 S3. 混合人工神经网络的训练混淆矩阵。

S2. 纯数字神经网络的设计

纯数字人工神经网络采用与混合神经网络的数字后端相同的顺序架构设计，包含四层：一个$N \times N$ 平均池化层；两个包含256个单元并使用 ReLU 激活函数的全连接神经元层；以及一个包含10个单元并使用 softmax 激活函数的全连接层。神经网络的优化采用完全相同的流程，包括优化器方案和迭代次数，并使用相同的硬件进行。优化流程包括使用 Adam 优化器、学习率为 0.001 的 150 个训练周期。大多数优化在~50 个训练周期内收敛。训练和验证准确率见图 S4。训练混淆矩阵见图 S5。

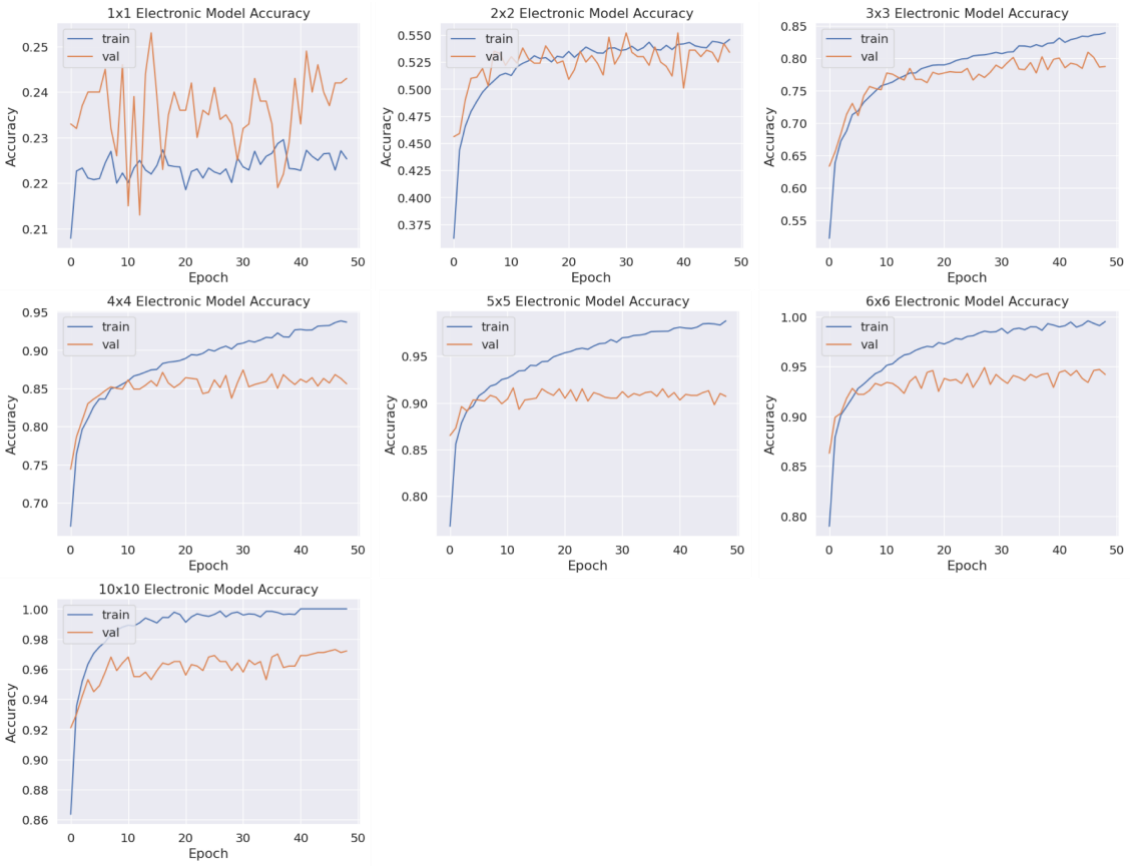
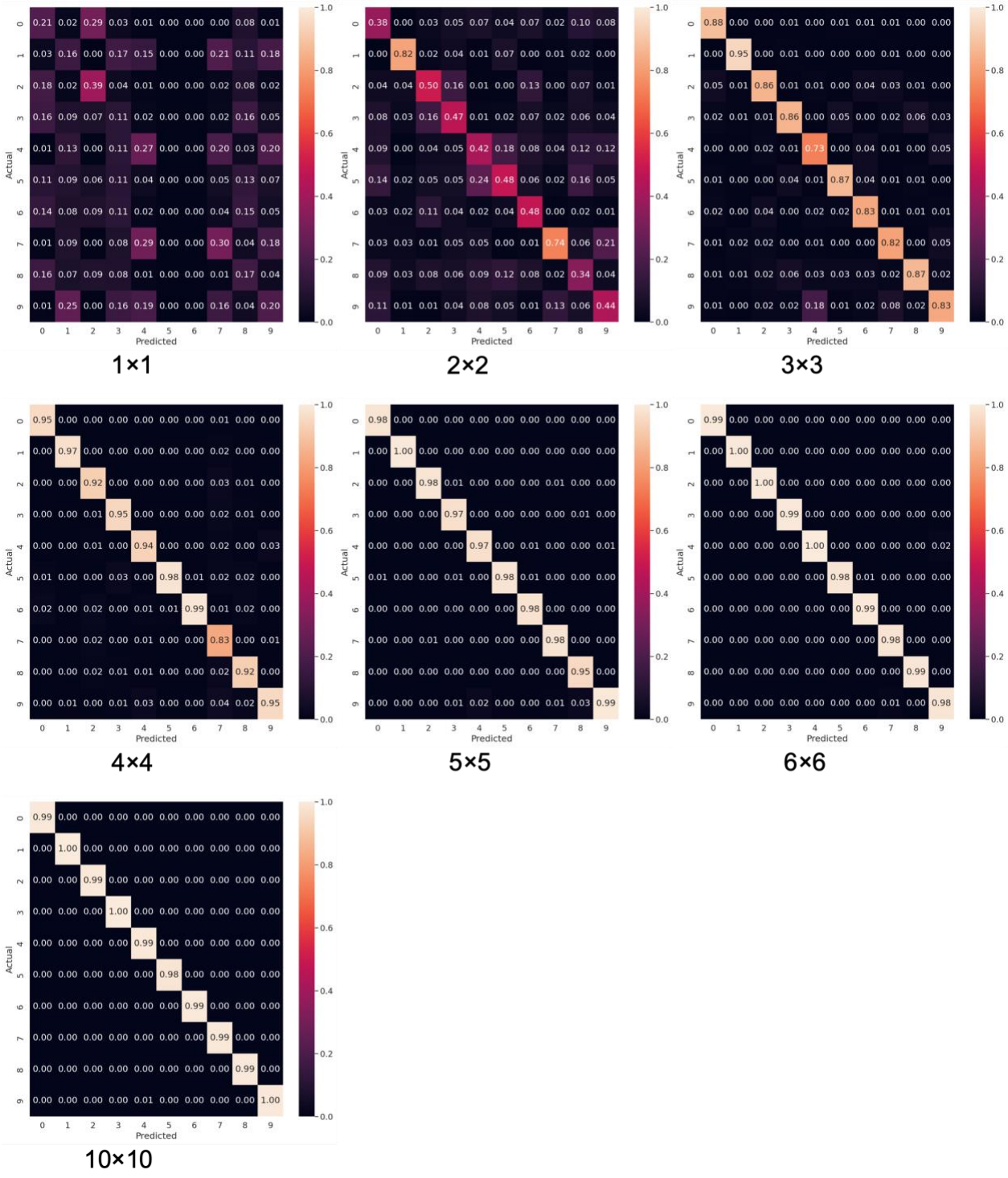


图 S4. 纯电子神经网络的训练（train）和验证（val）准确率。

我们注意到，为确保网络得到充分训练，我们从 ~5 层开始，逐步减少层数和神经元数量，并使用许多不同的输入进行训练。最终，仅使用两个隐藏层，我们就达到了较高的分类准确率（~97%）。我们强调，充分训练纯数字人工神经网络至关重要，否则我们可能会对光子学优势得出错误结论。我们怀疑，许多已发表的光学神经网络研究将其结果与训练不佳的数字人工神经网络进行了比较，从而显示出分类准确率有所提高。



Fi图 S5. 纯数字神经网络的训练混淆矩阵。

S3. 超构光学设计

该超表面由位于曼哈顿网格上的 SiN 超原子二维阵列组成，x 和 y 方向的周期均为 350 nm。每个超原子均呈方柱形，高度固定为 775 nm，横向宽度经过参数化设置，以覆盖网格周期范围。采用严格耦合波分析对该结构进行仿真[1], ，并将超表面的相位调制映射至能够产生最接近目标相位调制的相应超原子ϕ（x、 y）。这些超原子位于厚度约为500μm的石英晶圆上。

S4. 超构光学制备

在制备过程中，我们首先使用等离子体增强化学气相沉积（PECVD），在 SPTS PECVD 腔室中于厚度为 500 μm 的石英晶圆上沉积一层厚度为 ~ 775 nm 的 SiN 薄膜。随后将正性光刻胶（ZEP 520A）旋涂于晶圆表面，并在 180 °C 下烘烤 3 分钟。为最大限度地减小图形化过程中的充电影响，之后在其上旋涂一层导电聚合物（DisCharge H2O）。然后使用 100 kV 电子束（JEOL JBX6300FS）以 ~ 300 μC cm^{-2} 的剂量对光刻胶层进行图形化，并在乙酸戊酯中显影 2 分钟。接着，使用电子束蒸发沉积一层厚度为（~ 80 nm）的 AlO_x。在 90°C 加热的 NMP 中过夜剥离后，使用 C_4F_8/SF_6 混合气体，在电感耦合反应离子刻蚀机（Oxford PlasmaLab System 100）中将 SiN 层刻蚀至 700 nm（+/-2 nm）的深度，同时保留厚度为 ~10 nm 的 AlOx。为进行 SEM 成像，沉积了一层薄的导电 Au/Pd 层，以防止充电。

S5. 元光学编码器实验细节

对于混合神经网络实验测量，我们首先测量了标签-特征对。MNIST 特征，即手写黑白图像，使用最近邻插值放大后显示在 OLED 显示器（SmallHD 5.5英寸 Focus OLED HDMI显示器）上，如图 S6 所示。图像首先通过 OLED 显示器显示，该显示器与元光学编码器相距 ~10cm。随后，信号经过定制显微镜，将元光学编码器的输出传输到传感器。OLED显示屏的功耗为 ~15 W。不过，由于此处非相干光源的唯一用途是模拟真实世界中的物体，分类系统的实际功耗预算不应包括 OLED 显示器。

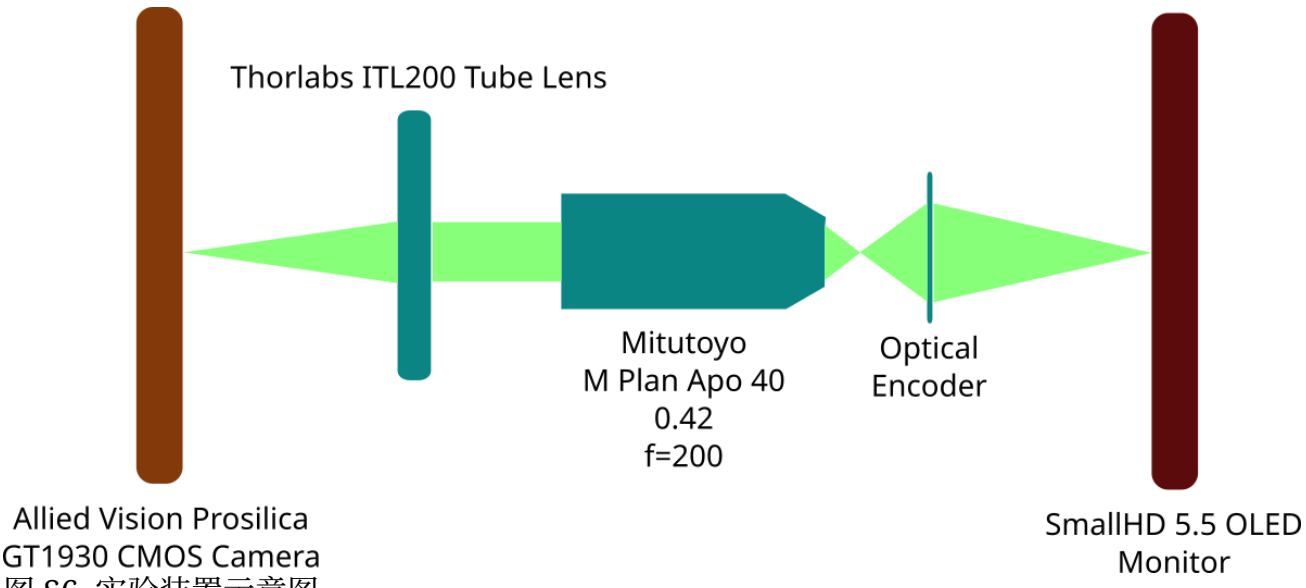


图 S6. 实验装置示意图。

图 S8 展示了部分实验捕获结果。首先，输入显示在顶行所示的 OLED 显示器上。随后，编码器处理输入并将信号投射到传感器上，如第二行所示。第三行展示了经过平均池化后的编码器输出。第四行展示了没有光学编码器时，纯数字人工神经网络接收到的内容。

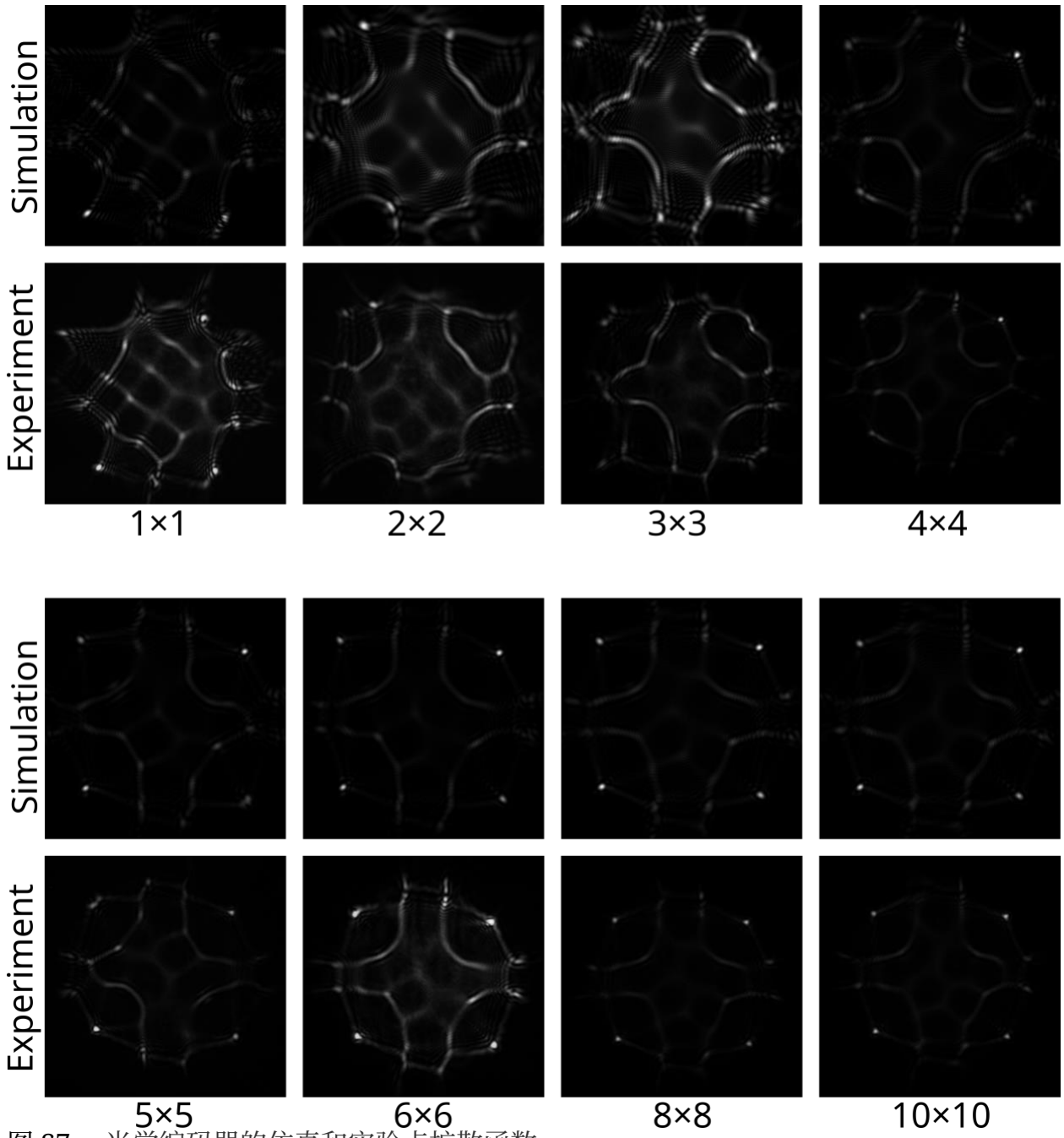


图 S7。 光学编码器的仿真和实验点扩散函数。

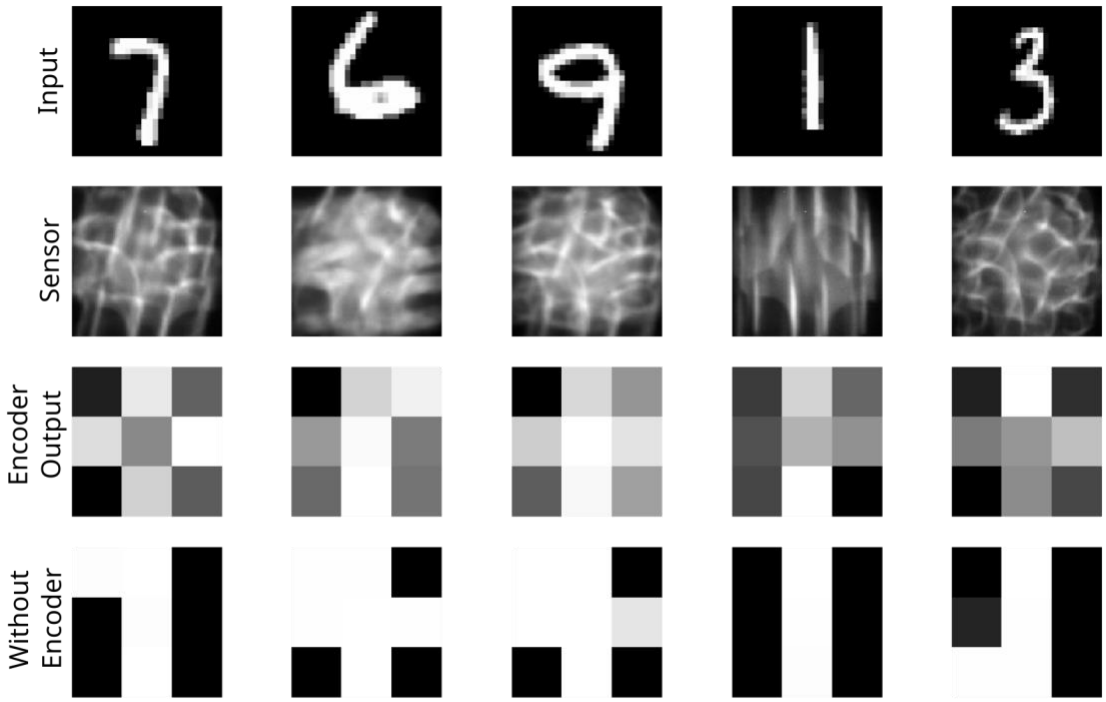


图 S8。 显示数字、捕获结果以及以下设备的输出：3 × 3 编码器。

对于每个元光学编码器，我们首先测量了点扩散函数响应，其与模拟点扩散函数吻合良好，如图 S7 所示。为了校准 OLED 显示器上的图像高度，我们显示了缩放因子各不相同的十字校准图案，并选择了与模拟输出最为相似的图像。随后，我们对图像进行适当裁剪，以反映仿真的正确物理范围。接着，使用平均池化将记录的图像降采样为 $N \times N$ 图像，并输入数字神经网络。值得注意的是，也可以通过使用较大的像素尺寸以光学方式实现平均池化步骤。较大的像素尺寸可能有利于低照度操作，这对于中波/长波红外波段的应用非常重要。

我们为训练集采集了 10,000 张图像，并为验证采集了 2,000 张图像。我们使用采集的验证数据集，首先计算理论混合神经网络的准确率，其准确率如图3(b)所示。随后使用训练数据集训练硬件在环混合人工神经网络，其拓扑结构与原始混合人工神经网络相同。在训练过程中，使用不同的随机种子重复此过程 4 次，并记录各次的验证准确率。我们还计算了验证准确率的标准差，并以正文图3(b)中误差棒之间的距离表示。

S6. 现有工作中光学优势的展示

目前，关于光学辅助神经网络的研究已有大量成果，但我们认为，尚无任何研究表明它们优于纯数字电子人工神经网络。本文所说的数字电子人工神经网络，包括纯软件方案（运行于图形处理器上）或基于专用集成电路（ASIC）的加速器。本质上，只要使用基于CMOS 晶体管的硬件，我们就将其归类为数字后端。任何纳米电子方案（例如基于忆阻器的方案）均不在本分析范围内。根据现有研究，光学神经网络大致可分为以下三类。

集成光子学 + 数字：大量研究依赖于基于集成光子学的光学运算和数字后端 [2]–[7]。其中大多数依赖某种形式的开关阵列，由马赫-曾德尔干涉仪或环形谐振器构成。这些网络大多受空间带宽积限制，而空间带宽积等同于波导数量。这本质上源于维度不足：在集成光子学中，我们实际上只有一个空间维度，因此空间带宽积为$N \sim A/\lambda$，其中 A 是芯片的维度，λ 是光波长。尽管受技术限制，目前波导数量要小得多，但集成光子学从根本上受限于可实现的空间带宽积。即使采用波分复用（WDM），输入向量的维度仍然较小。因此，WDM只能使 N线性扩展。空间带宽积有限，因而需要通过时域复用分批发送数据，这也要求在后端合并数据。因此，这些研究大多没有报告复用控制电路所带来的额外电功耗和延迟。此外，部分已解决的问题没有电子学基准，因此尚不清楚系统层面是否取得了任何优势。

大多数研究基本上表明，其分类准确率与数字人工神经网络相近，但并未计算专用数字集成电路的功耗和延迟。自由空间 + 数字：借助自由空间光学（空间光调制器或数字微镜器件），研究人员能够实现大得多的空间带宽积 [8]–[14]。得益于二维结构，空间带宽积按$N \sim (\frac{A}{\lambda})_2$扩展，其中 A 是光学系统的孔径，λ 是光波长。这种扩展方式远优于集成光子学中的线性扩展（即使采用WDM）。但大多数已发表的研究忽略了光学与电子学之间转换所带来的功耗和延迟。此外，使用空间光调制器还会增加大量功耗。将这些功耗数据与图形处理器进行比较也不公平，因为图形处理器旨在解决许多不同的问题，因此存在大量冗余。对于特定问题，可以通过剪枝/同或运算优化并设计专用集成电路，从而将功耗降得更低。

全光：最后，已有少数全光神经网络的演示（包括自由空间和集成光子学实现[3], [8], [10], [15]），无需数字后端即可实现。其中一些完全是线性的（不具备任何非线性），因此无法用于复杂数据集。此外，即使在这些演示中，通常也会使用多个光学层，每层都会带来额外的散射损耗。因此，输入功率可能较高，而且还需要激光器对电子信息进行编码。目前尚不清楚这些情况下实现了何种墙插效率，因此并未报告实际功耗。另一些研究则采用了基于可饱和吸收、冷原子中的电磁诱导透明或光电非线性的非线性机制（像增强器或光诱导电光非线性）。虽然其中一些研究确实演示了对MNIST数据集的运算，但并未核算所有可能来源的功耗，例如，未计入用于冷却

原子的激光功率。因此，所有这些研究都缺少对系统级功耗/延迟的真实估算，也就未能证实其相较于纯数字人工神经网络的优势。

参考文献[1]

V. Liu 和 S. Fan，“S4：一种用于分层周期结构的自由电磁求解器”，Comput. Phys. Commun.，第 183 卷，第 10 期，第 2233–2244 页，2012 年 10 月，doi: 10.1016/j.cpc.2012.04.026。[2] X. Xu 等，“用于光学神经网络的 11 TOPS 光子卷积加速器”，《自然》，第 589 卷，第 7840 期，第 44–51 页，2021 年 1 月，doi: 10.1038/s41586-020-03063-0。[3] F. Ashtiani、A. J. Geers 和 F. Aflatouni，“用于图像分类的片上光子深度神经网络”，《自然》，第 606 卷，第 7914 期，第 501–506 页，2022 年 6 月，doi: 10.1038/s41586- 022-04714-0。[4] Y. Shen 等，“利用相干纳米光子电路进行深度学习”，Nat. Photonics，第 11 卷，第 7 期，第 441–446 页，2017 年 7 月，doi: 10.1038/nphoton.2017.93。[5] A. Sludds 等，“互联网边缘的分布式光子深度学习”，Science，第 378 卷，第 6617 期，第 270–276 页，2022 年 10 月，doi: 10.1126/science.abq8271。[6] J. Feldmann 等，“使用集成光子张量核心进行并行卷积处理”，《自然》，第 589 卷，第 7840 期，第 52–58 页，2021 年 1 月，doi: 10.1038/s41586-020-03070-1。[7] S. Bandyopadhyay 等，“采用加速训练的单芯片光子深度神经网络”。arXiv，2022 年 8 月 2 日。访问日期：2023 年 4 月 4 日。[在线]。可访问：http://arxiv.org/abs/2208.01623[8] A. Ryou 等，“基于热原子非线性的自由空间光学神经网络”，Photonics Res.，第 9 卷，第 4 期，第 B128 页，2021 年 4 月，doi: 10.1364/PRJ.415964。[9] T. Wang、S.-Y. Ma、L. G. Wright、T. Onodera、B. C. Richard 和 P. L. McMahon，“每次乘法少于1个光子的光学神经网络”，Nat. Commun.，第 13 卷，第 1 期，第 123 页，2022 年 1 月，doi: 10.1038/s41467-021-27774-8。[10] T. Wang 等，“利用多层非线性光学神经网络进行图像传感”，Nat. Photonics，2023 年 3 月，doi: 10.1038/s41566-023-01170-8。[11] J. Chang、V. Sitzmann、X. Dun、W. Heidrich 和 G. Wetzstein，“采用优化衍射光学元件进行图像分类的混合式光电卷积神经网络”，Sci. Rep.，第 8 卷，第 1 期，第 12324 页，2018 年 8 月，doi: 10.1038/s41598-018-30619-y。[12] J. Spall、X. Guo 和 A. I. Lvovsky，“光学神经网络的混合式训练”，Optica，第 9 卷，第 7 期，第 803 页，2022 年 7 月，doi: 10.1364/OPTICA.456108。[13] D. Mengu、Y. Luo、Y. Rivenson 和 A. Ozcan，“ 衍射光学神经网络及其与电子神经网络集成的分析 ”，IEEE J. Sel. Top. Quantum Electron.，第 26 卷，第 1 期，第 1–14 页，2020 年 1 月，doi: 10.1109/JSTQE.2019.2921376。[14] H. Zhang 等，“用于实现复值神经网络的光学神经芯片”，Nat. Commun.，第 12 卷，第 1 期，第 457 页，2021 年 1 月，doi: 10.1038/s41467-020-20719-7。[15] X. Lin 等，“利用衍射深度神经网络进行全光机器学习”，Science，第 361 卷，第 6406 期，第 1004–1008 页，2018 年 9 月，doi: 10.1126/science.aat8084。