光子脉冲哈希检索的硬件实现

SHANGXUAN SHI、1 SHUIYING XIANG、1、2* XINTAO ZENG、1 YONGHANG CHEN、1 WANTING YU、1 YAHUI ZHANG1、2 XINGXING GUO、1 YUE HAO、2

[1]综合业务网国家重点实验室，西安电子科技大学，中国西安 710071[2]宽禁带半导体技术国家重点学科实验室，微电子学院，西安电子科技大学，中国西安 710071 *通信作者：syxiang@xidian.edu.cn

接收于 XXXX 年 XX 月X；修订于 XXXX 年 XX 月；接受于 XXXX 年 XX 月；发布于 XXXX 年 XX 月（文档编号 XXXXX）；发表于 XXX 年 XX 月 X

哈希检索是大规模相似性搜索的关键技术，广泛应用于大语言模型（LLM）的检索增强生成（RAG）、海量图像库以及生物信息学序列匹配。然而，传统电子哈希实现处理高维数据时，在功耗和延迟方面面临严重瓶颈，而现有光子神经网络通常缺乏在模拟噪声下直接生成稳健二进制码的机制。为应对这些挑战，我们提出了一种软硬件协同设计的光子脉冲哈希框架。我们利用带可饱和吸收体的分布反馈激光器（DFB-SA）的非线性阈值动力学，实现单步脉冲神经网络（SNN）的最终二值化。尤为关键的是，我们引入了硬件感知量化间隔损失，以最大化决策间隔，有效缓解由光强波动引起的位翻转。在 MNIST（图像）和 20 Newsgroups（文本）数据集上的验证表明，我们的系统能够稳健地生成二进制码，并实现与数字基线相当的高检索准确率。更重要的是，所提出的光子架构展现出卓越的效率，编码延迟为 2.294 ns/查询，能耗为 73.70 pJ/查询。该工作为高吞吐量信息检索任务中的超快速、低能耗光电神经形态计算提供了一条稳健且可行的路径。

http://dx.doi.org/10.1364/AO.99.099999

1. 引言

相似性搜索技术是信息时代的基石，在许多前沿应用中发挥着关键作用。例如，在大语言模型（LLM）推理中，检索增强生成（RAG）利用外部知识源来提高事实准确性和可靠性 [1]。在海量多媒体库中，高效检索是搜索引擎和推荐系统的基础，能够快速找到与查询相似的项目 [2,3]。在生物信息学中，对大规模基因组序列（DNA/RNA）进行同源性搜索，本质上是一个高维相似性匹配问题，是疾病诊断和药物发现的核心环节 [4,5]。因此，在保持能效的同时实现高检索吞吐量，已成为这些领域的一项关键要求。

为应对大数据时代大规模数据检索所面临的挑战，学术界广泛推动了基于哈希的近似最近邻搜索方法的发展。与代价高昂的精确搜索策略不同，这些方法将高维多媒体数据投影到低维汉明空间中的紧凑二进制编码，在保留语义相似性的同时，大幅降低存储需求和计算开销 [6,7]。以往的解决方案主要依赖数据无关投影或手工设计特征 [8],，而深度哈希的出现通过采用深度神经网络进行端到端特征提取和二进制编码学习，彻底改变了这一领域。这种范式转变能够捕捉浅层方法往往无法发现的复杂非线性语义结构。代表性架构包括深度哈希网络（DHN）[9],、深度监督哈希（DSH）[10],和深度成对监督哈希（DPSH）[11]。值得注意的是，HashNet[12] 通过采用延续优化策略解决符号函数的梯度不稳定性问题，进一步改善了收敛性。然而，这些方法大多在电子计算平台上实现，实际部署时受到冯·诺依曼瓶颈的严重制约；具体而言，有限的互连带宽和不断增长的功耗，使得同时实现超大规模并行性与低延迟响应面临困难 [13,14]。因此，光子计算日益受到关注。它利用光子信号进行信息传输和处理，具备高吞吐量并行性、宽带宽以及有望提升能效等固有优势 [15,16]。

光子计算的最新进展在计算机视觉和时序信号处理等领域展现出巨大潜力。在计算机视觉应用中，研究人员已成功构建片上光学卷积神经网络，用于图像分类 [17,18]。光子储备池计算和循环神经网络也已用于时间序列预测和信道均衡等任务 [19,20]。近期，光学 Transformer 架构和非线性

激活函数也已有相关报道 [21–23]。在这些研究方向中，光子脉冲神经网络 （SNN）尤其受到关注，因为基于脉冲的信息表示与光脉冲动力学天然契合。通过模仿生物神经行为，SNN 利用时空脉冲进行事件驱动计算，并在特定条件下相比传统人工神经网络实现更高的能效 [24,25]。研究人员探索了多种物理机制，以实现高性能光子脉冲神经网络，涵盖相变材料 [26] 、石墨烯覆硅架构 [27] 以及利用稀疏性的大规模网络 [28]。值得注意的是，Xiang 等通过利用非相干光注入下带可饱和吸收体的分布式反馈激光器（DFB-SA）的非线性动力学，实现了光子神经形态模式识别 [29]。

尽管已有这些进展，专门面向基于哈希的相似性检索的光子脉冲神经网络架构在很大程度上仍未得到探索。现有光子脉冲神经网络演示通常产生针对分类优化的模拟脉冲相关信号，而检索应用需要严格的 0/1 二进制编码，以实现高效的汉明距离搜索。因此，一个关键挑战是将光子脉冲动力学映射为稳健的二进制哈希码，尤其是因为模拟光子器件对噪声敏感。

本文提出了一种面向相似性检索的软硬件协同设计的光子脉冲哈希框架。主要贡献如下。(1) 我们提出了一种面向检索的脉冲哈希框架，适用于单步脉冲设置，通过物理阈值判决将输入特征映射为二进制哈希码，从而在算法表示与器件行为之间建立直接对应关系。(2) 为缓解噪声导致的基于阈值的光子二值化中的位翻转，我们设计了一种硬件感知量化间隔损失，明确促使连续输出远离阈值不确定区域，从而提升二值化稳定性以及器件噪声下的检索鲁棒性。(3) 我们在图像和文本模态上验证了所提出的框架，并进一步将其与基于 DFB-SA 的脉冲阈值行为相结合。我们通过实验验证了 DFB-SA 激光器生成二进制脉冲的能力，并展示了其与所提出的哈希层的兼容性，为光子神经形态检索提供了一条可行的高性能实现路径。

2. 总体架构与方法

A. 系统架构概述

图 1(a) 展示了所提出的软硬件协同设计的光子脉冲哈希框架，该框架由四个协同的功能模块构成。流程始于多模态特征提取和 SNN 编码。原始数据（如图像和文本）首先利用轻量级预处理技术压缩为低维特征向量，以适应光子硬件有限的输入带宽。随后，这些向量被输入单步 SNN，该网络执行单次前向传播，在最终层生成连续的膜电位。具体而言，这些连续值通过二值化进行编码：大于阈值的值编码为 1，小于或等于阈值的值编码为 0。

从连续势生成此类二进制哈希码时，一个关键挑战是它们容易受到模拟扰动的影响，

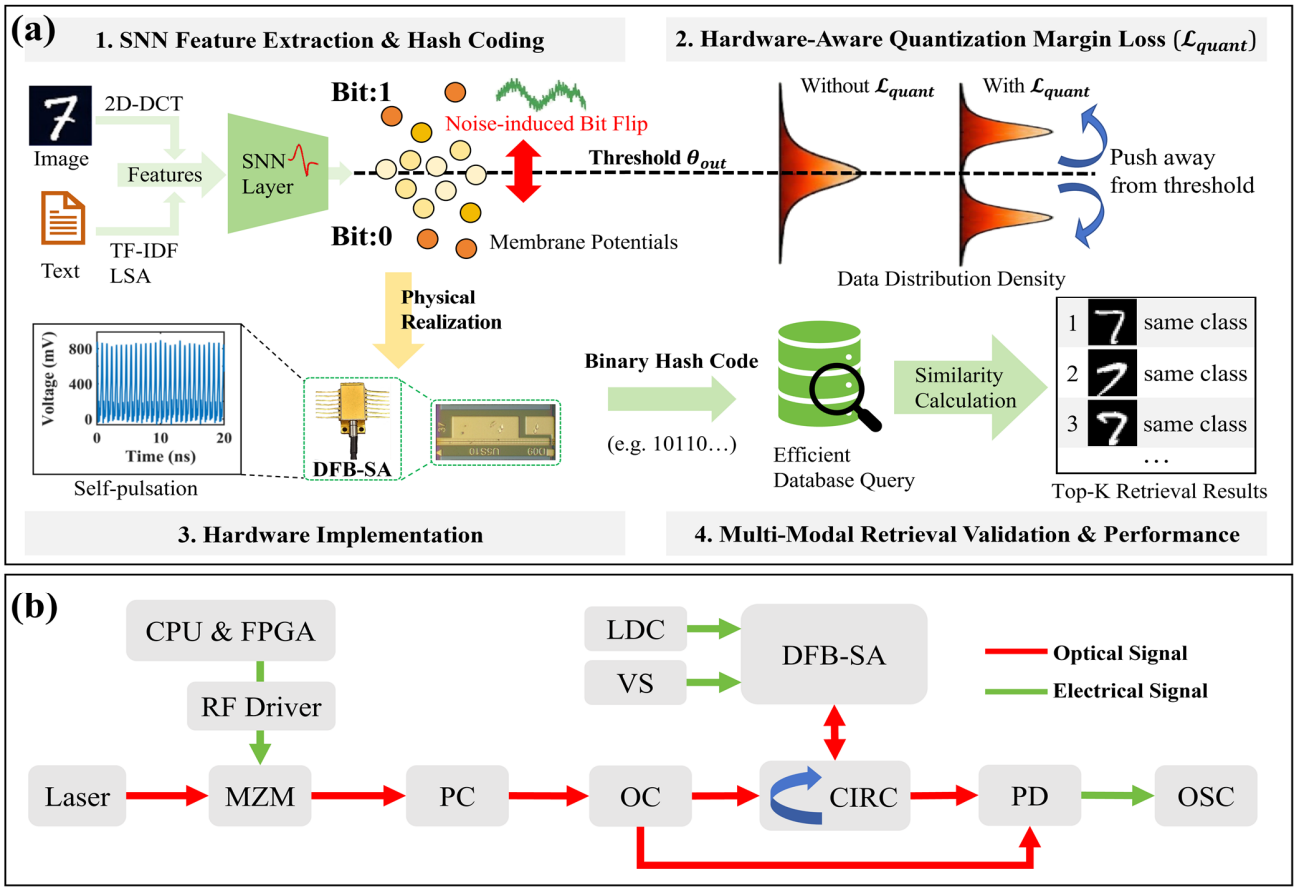

图 1. (a)　所提出的光子脉冲哈希检索框架示意图。(b) 基于 DFB-SA 激光器的硬件实现

r.

从而在决策边界附近引发噪声引起的比特翻转。为解决这一问题，该框架引入了硬件感知量化间隔损失（L_{quant}）。如数据分布密度比较所示，这种面向裕度的优化会明确推动预激活信号远离二值化阈值，从而有效消除阈值不确定区域，大幅增强0/1决策过程抵抗模拟硬件噪声的能力。

至关重要的是，算法硬阈值 T_{out} 与 DFB-SA 激光器的固有物理特性系统性地保持一致。连续的预激活值被用作驱动信号，直接映射到物理器件。通过利用 DFB-SA 激光器的脉冲响应，该框架将这些模拟输入转换为稳定的二进制脉冲物理实现。最后，这些由硬件生成的二进制哈希码被用于大规模数据库查询中的相似度计算，从而实现高效且准确的多模态检索。

B. 脉冲哈希检索算法

(1) 预处理与特征构建

为适应光子硬件有限的扇入容量，我们在 SNN 阶段之前实现了一个轻量级预处理模块。该模块将高维原始数据映射到低维特征空间，在满足设备输入要求的同时保留稳健的语义信息。

图像模态：对于输入的灰度图像，我们采用标准的二维离散余弦变换（2D-DCT） [30] 提取低频分量。由于图像的能量通常集中在低频区域 [31], ，这种常规方法能够有效压缩空间冗余，同时保留供后续网络使用的关键信息结构。

文本模态：对于文本语料库，我们采用常规的自然语言处理流程。首先，构建一个按词频-逆文档频率（TF- IDF）加权的词项-文档矩阵 [32]。为去除稀疏噪声并将维度压缩至适合光子链路的规模，我们通过截断奇异值分解（SVD）进行潜在语义分析（LSA）[33]。此步骤将文档投影至稠密的低维潜在语义空间，从而显著降低后续 SNN脉冲哈希编码的输入开销。

*(2) SNN-Hash*网络与脉冲哈希编码

在本研究中，我们将 SNN-Hash 实现为一个堆叠全连接网络。该网络由线性映射层与脉冲激活层交替组成。我们采用单步SNN策略，实现针对硬件效率优化的确定性一次性编码过程。具体而言，单步操作避免了多周期时间积分和重复的设备输入/输出，从而降低延迟和系统复杂度。此外，它还产生一种直接的计算模式：先进行线性映射，再执行阈值激活。这种模式与光子矩阵-向量乘法及 DFB-SA 激光器的脉冲响应天然匹配，因此输出脉冲可直接解释为哈希比特。

将网络最后一层的连续预激活输出记为$\check{z}_{out} \in R^K$，其中$\check{z}$为哈希码长度（即比特数）。最终的二进制哈希码$b_{out} \in \{0,1\}^K$通过固定阈值生成：

$$b_{out} = \mathcal{H}(z_{out} - \theta_{out}),\ \theta_{out} = 1.0 \qquad \textbf{(1)}$$

其中H(·)为逐元素应用的海维赛德阶跃函数，用于输出二进制脉冲：

$$\mathcal{H}(a) = \begin{cases} 1, & a > 0, \\ 0, & a \le 0. \end{cases} \qquad \textbf{(2)}$$

此处，T_{out} 被固定，以与DFB-SA激光器建立硬件对齐接口。在系统层面，该设计将网络的二值化规则映射到DFB-SA激光器的脉冲响应，为后续的软硬件协同验证和闭环实验提供直接基础。

*（3）*损失函数与优化

损失函数设计

为了训练光子脉冲神经网络生成鲁棒的二进制哈希码，同时遵循基于阈值的硬件接口，我们采用复合目标函数。总损失定义如下：

$$\mathcal{L} = \mathcal{L}_{pair} + \lambda_{bal}\mathcal{L}_{balance} + \lambda_{ind}\mathcal{L}_{indep} + \lambda_{quant}\mathcal{L}_{quant} \qquad \textbf{(3)}$$

其中，η_{bal}、η_{ind}和η_{quant} 为超参数权重。L_{pair}、$L_{balance}$ 和 L_{indep} 是改编自 HashNet [12]的标准哈希目标函数。具体而言，加权成对相似性损失（L_{pair}）利用双极性编码（±1）的内积，使相似样本对彼此靠近，并使不相似样本对在汉明空间中彼此远离。比特平衡损失（$L_{balance}$）通过促使每个比特取 0 和 1 的概率相等来最大化信息容量；而比特独立性损失（L_{indep}）则通过最小化非对角相关性来减少哈希比特间的冗余。

光子计算中的一项关键挑战是阈值化前的模拟噪声（例如光强/功率波动）。对于大小为 ∞ 的小批量，令 $\check{z}_{out,i} \in R^K$ 表示样本 i 在哈希层的阈值前输出向量（连续驱动信号），并令 $T_{out} = 1.0$ 表示与硬件对齐的固定阈值。如果第 k 位分量 $\check{z}_{out,i,k}$ 过于接近 T_{out}，微小扰动就可能改变二值化判定并导致位翻转。我们不再最小化传统的量化误差（例如 $\| b - \check{z} \|$），而是引入一种基于间隔的损失，使预激活值远离阈值：

$$\mathcal{L}_{quant} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K} exp\left(-\alpha\left|z_{out,i,k} - \theta_{out}\right|\right) \qquad \textbf{(4)}$$

其中 $\dagger > 0$ 控制陡峭程度。该项会惩罚接近 DFB-SA 开关边界的输出，减少落入器件阈值不确定区域的样本，从而提高 0/1 判定在.....条件下的鲁棒性

p光子噪声。

两阶段阈值校准

在单步脉冲神经网络的训练中，一个常见的性能退化问题是隐藏层神经元过度稀疏，甚至出现“静默神经元”（脉冲全为零），这会在反向传播过程中显著削弱梯度和学习信号 [34]。为缓解这一问题，我们在正式训练前引入两阶段阈值校准策略，确保初始化阶段达到适当的脉冲稀疏度。该策略不依赖任意初始值，而是根据预先设定的目标发放率动态调整阈值。通过观察小规模校准子集上的预激活分布（不更新权重），我们采用基于分位数的方法依次校准各隐藏层。这种逐层初始化可避免早期出现静默脉冲，从而在实际训练阶段获得稳定的梯度并改善收敛。

光子脉冲哈希检索框架的整体训练流程以及该阈值校准的详细分步过程，总结于 **算法 1**。

算法 1：光子脉冲哈希

检索的训练流程

需要： 训练集 $\mathcal{D}_{tr} = (ب_i, ب_i)$，测试/查询集 $\mathcal{D}_{te}$；
预处理算子 $\mathcal{P}r(\cdot)$（图像：离散余弦变换；文本：词频-逆文档频率+潜在语义分析）；哈希长度 ك；脉冲神经网络模型 $f_\theta(\cdot)$；固定输出阈值 $T_{out} = 1.0$；目标发放率 p_{fire}；训练轮数 ة，批量大小 ه；损失权重 η_{bal}, η_{ind}, η_{quant}.

确保： 最佳模型参数 θ^*。

预处理：
$F_{tr} \leftarrow P\text{ك}(\mathcal{D}_{tr})$, $F_{te} \leftarrow P\text{ك}(\mathcal{D}_{te})$；使用训练统计量进行归一化。初始化 f_θ（线性权重和隐藏层脉冲阈值）。

两阶段阈值校准：
对每个隐藏层 $l \leftarrow 1$ 到 $\text{ي} - 1$ 执行
1. 从 F_{tr}中采样校准子集 B_{cal}。2. 经第 $1, \ldots, l$ 层前向传播，以在 B 上收集预激活 $\text{ب}^{(l)}_{cal}$。3. 根据观测到的 $\text{ب}^{(l)}$ 分布的第（$1 - p_{fire}$）分位数，逐元素更新阈值 $T^{(l)}$。4. 固定 $T^{(l)}$ 并进入下一层。

结束循环

主训练循环：
对 epoch $e \leftarrow 1$ 至 ة 执行
1. 采样一个类别均衡小批次 $\{(ب_i, ب_i)\}_{i=1}^{B}$。2. 通过隐藏层进行前向传播，计算输出预激活值 $ب_{out}$。3. 生成二进制哈希码 $b_i = H(z_{out,i} - T_{out})$ 和双极表示 $ب_i = 2b_i - 1$。4. 使用 $ب_i$、$ب_{out}$ 和 $ب_i$ 计算复合损失 $L \leftarrow L_{pair} + \eta_{bal}L_{bal} + \eta_{ind}L_{ind} + \eta_{quant}L_{quant}$。5. 通过带有替代梯度的反向传播更新 θ。结束循环，返回 θ^*

C. 硬件—软件协同计算实验设置

图 1(b) 展示了使用 DFB-SA 激光器实现硬件对齐阈值操作的实验装置。连续波激光器提供光载波，该光载波由马赫–曾德尔调制器（MZM）进行强度调制，调制器由高速电信号驱动。驱动波形由 CPU/FPGA平台生成，并通过射频（RF）驱动器传输，以确保足够的带宽和振幅。随后，调制光信号通过偏振控制器（PC），使其偏振态与最佳耦合所需的状态相匹配。偏振对准后，信号经由光耦合器（OC）路由，并通过三端口光环形器（CIRC）注入 DFB-SA 激光器。

DFB-SA激光器用于生成类阈值光尖峰：当注入的驱动电平超过器件的有效可激发阈值时，会产生一个“1”尖峰；否则，输出保持为“0”。随后，光电探测器（PD）检测生成的光学输出，并使用实时示波器（OSC）记录，以便采集波形并进行后续比特提取。激光二极管控制器（LDC）为DFB-SA激光器的增益区提供低噪声偏置电流和温度稳定，而外部电压源（VS）则为可饱和吸收体区提供反向偏置电压。在图1(b)中，红色和绿色箭头分别表示光信号路径和电信号路径。该设置在软件生成的预激活值 $ب_{out}$ 与哈希层的硬件二元判决之间建立了直接接口，从而能够对硬件对齐的二进制哈希码生成进行闭环验证。

3. 性能评估与实验结果

为验证所提出的软硬件协同设计光子脉冲哈希检索框架的有效性，我们在标准图像和文本检索基准上进行了系统实验。我们首先介绍实验方案和

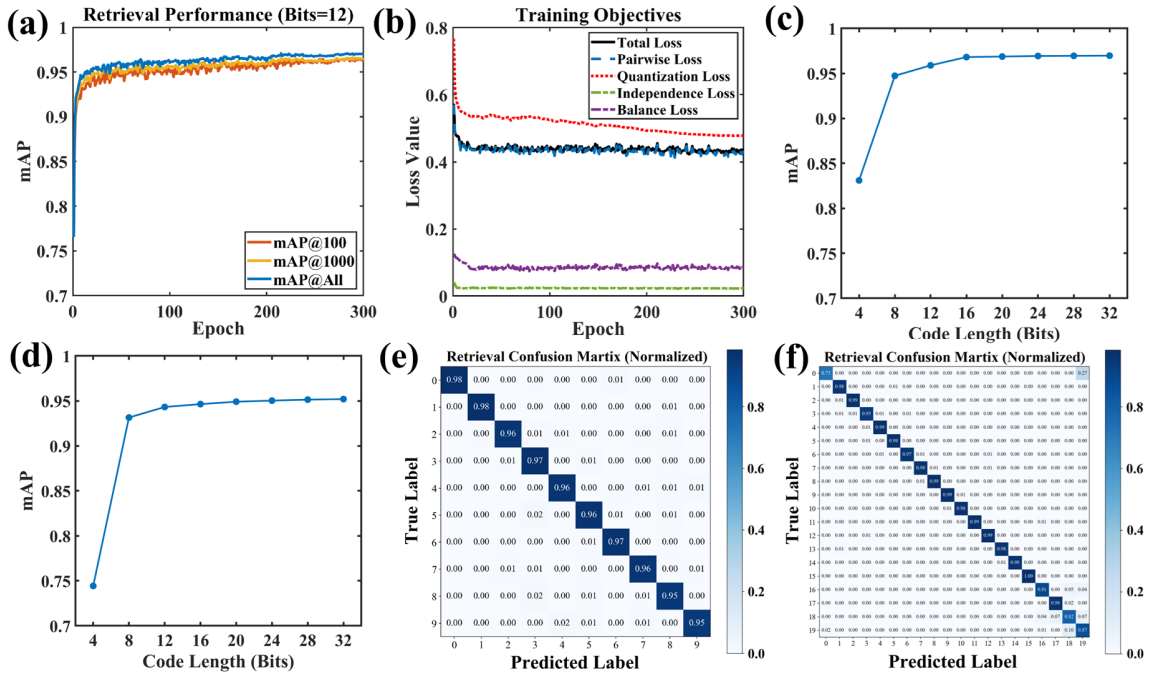

图 2. 训练动态与检索性能评估。(a) 在 12 位配置下，训练过程中平均准确率均值（mAP）的上升曲线，以及 (b) 训练损失的下降曲线。检索平均精度均值随哈希码长度变化的曲线，分别对应 (c) MNIST 和 (d) 20 Newsgroups（20NG）数据集。(e) 图像（MNIST）和 (f) 文本（20NG）基准测试的归一化检索混淆矩阵。

评估指标。随后，我们分析：（1）不同指标下的训练收敛情况；（2）不同哈希长度下的检索性能；（3）定性检索示例和特征空间组织；以及（4）DFB-SA 阈值判决前对模拟扰动的鲁棒性。

A. 实验设置与指标

（1）数据集

我们在两个广泛使用的基准上进行评估：MNIST（图像模态；10个类别共60,000个训练样本和10,000个测试样本）和20 Newsgroups（20NG，文本模态；约20,000篇文档，涵盖20个主题）。对于MNIST，图像被转换为紧凑的低频DCT特征（第2节）。对于20NG，我们构建TF-IDF特征，随后进行截断奇异值分解/潜在语义分析，以获得稠密表示（第2节）。

（2）模型配置与训练。

我们采用单步 SNN-HashNet。隐藏层使用具有可学习的逐神经元阈值的脉冲神经元通过替代梯度进行优化，而最终哈希层使用固定的硬阈值 T_{out}，该阈值与 DFB-SA激光器的开关激发阈值特性相一致（第 2 节）。除非另有说明，模型均使用 Adam 进行训练，学习率为 0.001。此外，在构建成对监督时，采用类别均衡小批量采样，以减小数据不平衡的影响。

（3）检索指标（平均准确率均值）

检索性能使用平均准确率均值（平均准确率均值）进行评估。设Q 表示包含$|Q|$ 个查询的查询集。对于查询i，我们按汉明距离对数据库条目进行排序，并计算前ㄴ个检索结果内的平均准确率：

$$mAP@K = \frac{1}{|Q|}\sum_{i=1}^{|Q|} AP_i(K), \quad (5)$$

$$AP_i(K) = \frac{1}{R_i(K)}\sum_{j=1}^{K} P_i(j)\cdot rel_i(j) \quad (6)$$

其中，$P_i(j)$ 表示查询 i 的前 j 个检索结果中的准确率，ㄴ $el_i(j) \in \{0,1\}$ 表示第 j 个检索结果是否相关（属于同一类别/主题），而 $R_i(ㄴ)$ 表示前 ㄴ 个结果中相关项的数量。我们报告

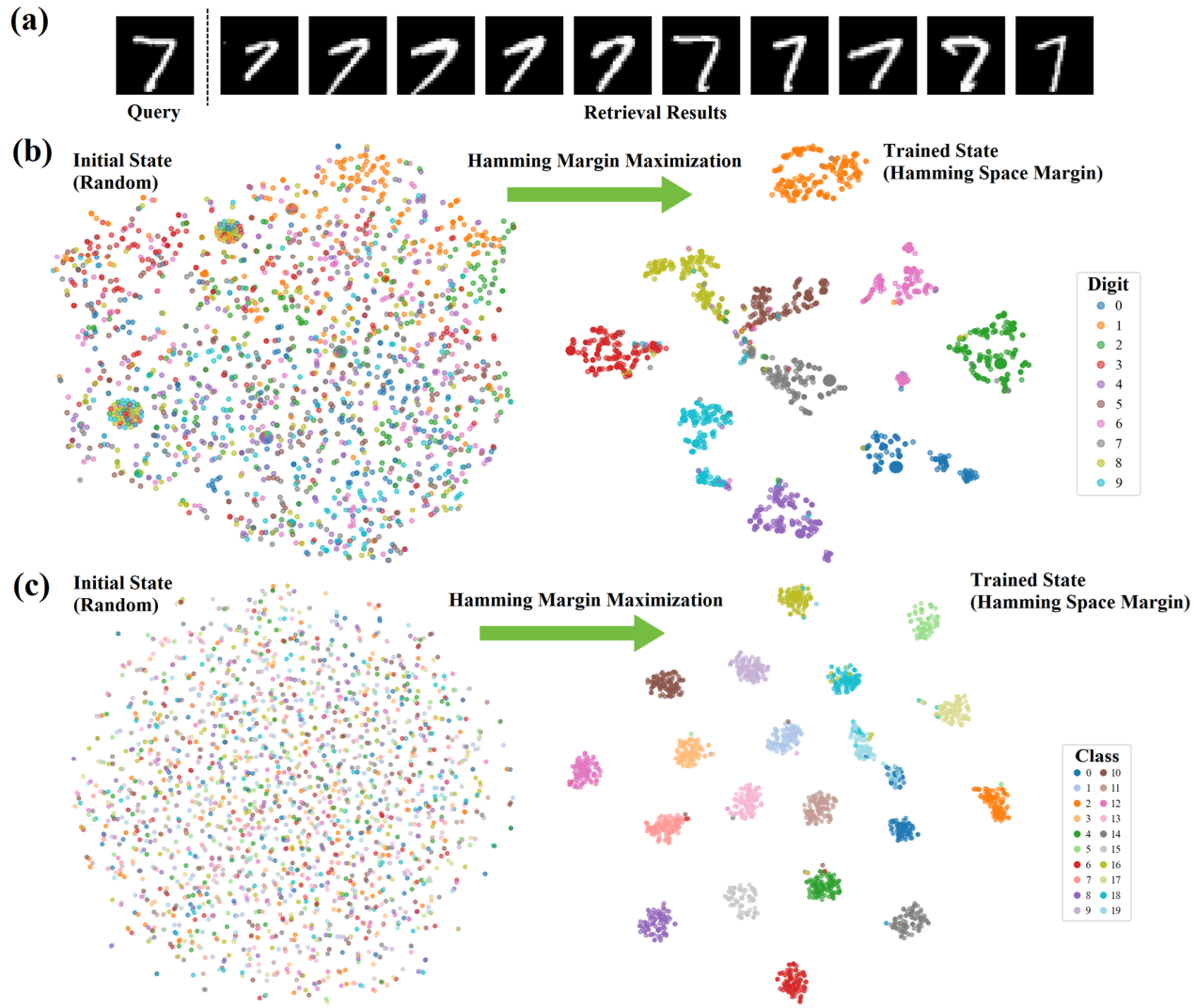

图 3. (a) 一个典型的 MNIST 检索示例。(b, c) 初始（随机）状态和训练状态下的特征空间分布，展示了间隔扩大和簇分离。

不同编码长度（位数）下的平均准确率均值，并且还报告考虑完整排序列表时的 mAP@All。

B. 12位哈希下的训练收敛

图2展示了所提出框架的训练动态和整体检索性能。聚焦于12位配置，图2(a)显示，检索准确率（mAP@100、mAP@1000 和 mAP@All）在训练初期迅速提升，随后逐渐趋于饱和，表明模型能够稳定地学习汉明空间相似度。同时，图2(b)表明，总目标函数持续下降并趋于稳定，证实了单步 SNN 架构下优化过程的可靠性。

观察到的收敛由复合目标函数驱动，该函数结合了成对相似性保持与编码正则化。成对项提供主要的判别信号，使相似样本对在汉明空间中彼此靠近，同时使不相似样本对彼此远离。同时，平衡正则化项和独立性正则化项可防止出现退化编码（例如编码坍塌或比特高度相关），从而最大化信息容量。值得注意的是，硬件感知量化间隔损失在整个训练过程中始终有效，主动促使连续的阈值前输出 $ب_{out}$ 远离固定阈值 T_{out}。这一特性对光子硬件部署尤其有利，因为接近 DFB-SA 可激发阈值边界的输出最容易受到噪声引起的比特翻转影响。

C. 检索性能与类别可区分性

图2(c)和图2(d)分别展示了 MNIST 和 20NG 上平均准确率均值随哈希长度的变化。在 MNIST 上，平均准确率均值在 4 位时约为 0.83，随着编码长度达到 16 位及以上，很快趋于饱和并超过 0.97，这表明中等长度的编码足以在低碰撞概率下编码数字级语义。在 20NG 上，平均准确率均值也呈现类似的饱和趋势，在较长的编码长度下（例如 ≥20 位）达到约 0.95 的稳定平台。这些趋势表明，增加编码长度可以提升表征能力并减少汉明碰撞；但当语义可分性成为主导因素后，性能增益会逐渐减小。

图 2(e) 和 2(f) 给出了根据检索结果得出的归一化混淆矩阵。对于 MNIST数据集，对角线数值始终高于 0.95，体现出较强的类别间一致性，以及汉明空间中的低碰撞率。对于 20NG数据集，该模型在大多数类别上都保持较高的准确率（通常高于 0.90），这表明，所提出的光子脉冲哈希流程在统一架构下能够很好地从视觉推广到文本。

D. 定性检索示例与特征空间组织

图3(a)展示了一个具有代表性的 MNIST 数字“7”检索示例。给定一张查询图像，系统能够从数据库中成功检索出最相似的前 10 个项目，且这些项目的语义标签均相同。这证实，光学阈值层生成的二进制编码保留了输入数据的基本语义相似性。图3(b)和图3(c)对比了初始（随机）状态与训练状态下的特征分布，揭示了这一性能背后的机制。初始时，不同类别的特征高度重叠，无法通过简单的阈值化加以区分。然而，经过复合损失函数训练，尤其是采用汉明间隔最大化后，特征明显分离，形成紧凑且彼此分隔良好的簇。对于图像和文本模态，类内距离均得到最小化，同时类间间隔显著扩大。这种“聚类效应”确保预激活值远离 DFB-SA 激光器的可激发阈值，从而增强光学输出对潜在硬件噪声或功率波动的鲁棒性。

E. 鲁棒性分析

在实际的光子硬件部署中，光子链路/器件中的模拟扰动——例如光功率/强度波动、耦合变化以及电学/光子读出噪声——会直接影响基于阈值的二值化稳定性。在我们的框架中，最终的哈希决策通过应用一个固定硬阈值来生成，该阈值与 DFB-SA 激光器的可激发阈值特性相匹配。令 $ب_{out} \in R^K$ 表示哈希层的连续阈值前输出，$T_{out} = 1.0$ 表示与硬件对齐的固定阈值。当 $ب_{out,\ k}$ 的任一分量接近 T_{out} 时，即使很小的扰动也可能改变二值决策并导致位翻转，进而累积造成汉明距离检索性能的显著下降。

为量化有多少样本在阈值边界附近运行，我们分析输出裕量 $ب_{out} - T_{out}$。如图4(a)所示，NO-QUANT基线的裕量分布在零附近呈现尖锐峰值，表明许多输出聚集在阈值附近，因此对扰动高度敏感。相比之下，引入量化裕度损失（QUANT(0.01) 和 QUANT(0.05)）会使裕量分布向两侧扩展，大幅减少阈值附近区域的概率质量

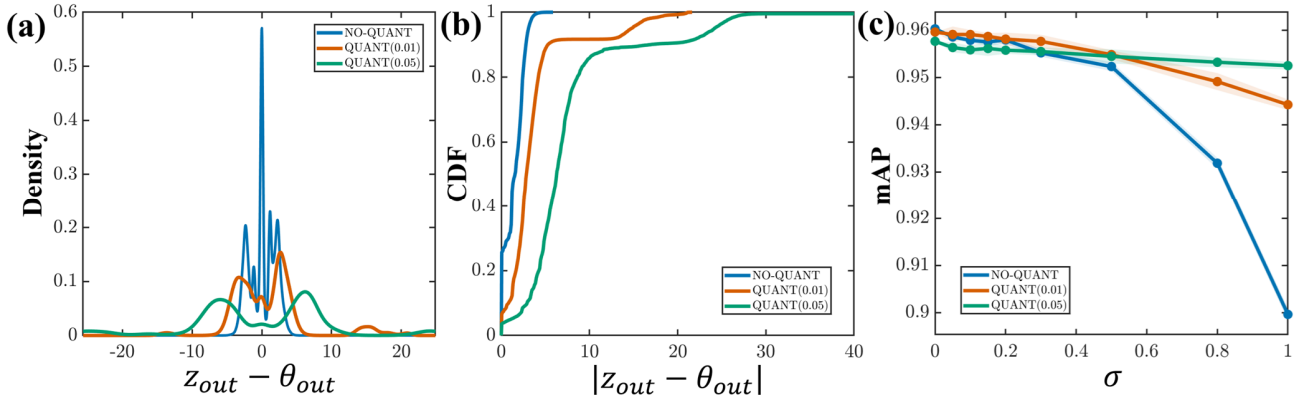

图4. 对模拟噪声的鲁棒性分析。(a) 输出裕量（$ب_{out} - T_{out}$）的密度分布和 (b) 累积分布函数，展示了所提出的量化损失所带来的安全裕度提升。(c) 检索平均精度均值随高斯噪声强度 σ 的变化，表明其对噪声引起的比特翻转具有更强的鲁棒性。

并在 T_{out} 周围形成更大的“安全裕度”。图4(b)中 $|z_{out} - T_{out}|$ 的累积分布函数（CDF）进一步证实了这一趋势：采用裕度正则化后，CDF 在裕量较小时上升得慢得多，这意味着落入阈值敏感区域的样本比例更小。

为模拟 DFB-SA 阈值判决前的模拟/读出扰动，我们在连续输出中注入零均值高斯噪声：

$$\tilde{z}_{out} = z_{out} + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 I) \quad (7)$$

其中，σ 控制噪声强度。随后，我们应用固定阈值 T_{out} 以获得二进制哈希码，并评估不同 σ 下的检索平均精度均值。

图 4(c) 显示，在无噪声或低噪声条件下（例如，$\sigma \leq 0.2$），三种设置的性能相近，表明裕度正则化不会损害名义检索准确率。随着 σ 增大（例如，$\sigma \geq 0.5$），NO-QUANT 模型的平均精度均值下降得快得多，而采用裕度损失训练的模型性能下降更为平缓，其中 QUANT(0.05) 表现出最强的鲁棒性。这些结果与其内在机制相符：通过显式增大 $|z_{out} - T_{out}|$，量化裕度损失降低了扰动使样本跨越阈值边界的概率，从而抑制噪声引起的比特翻转，并提升模型在基于 DFB-SA 的光子阈值处理下的可部署性。

4. 软硬件协同计算实验结果

为验证所提出的软硬件协同设计光子脉冲哈希框架的可行性，我们使用 DFB-SA 激光器作为物理阈值器件，在 MNIST 和 20NG 基准数据集上开展端到端检索实验。具体而言，我们从每个测试集中随机选取 100 个类别均衡的查询（MNIST：每类 10 个查询；20NG：每类 5 个查询）。为对这些查询进行物理阈值化激活，我们采用 DFB-SA 激光器，该激光器由电流驱动增益区和电压偏置可饱和吸收体（SA）区组成。正如我们先前的工作 [36], 中所述，优化后的 DFB-SA芯片工作于 1550 nm 通信波段，激射阈值低至约 28 mA。实验期间，器件温度严格稳定在 25°C，以确保类神经元动力学稳定。通过精细调节电子偏置（例如，将增益电流设为 32 mA，将 SA 反向电压设为 -1.8 V），激光器表现出类似生物神经元的非线性阈值响应。

对于每个查询，训练好的 SNN-HashNet 会生成连续的阈值前输出 z_{out}，并将其作为驱动信号输入 DFB-SA 阈值处理模块。利用其固有的阈值响应，DFB-SA 激光器将连续驱动振幅转换为稳定的 0/1 脉冲输出，形成最终的二进制哈希码 $b_{out} \in \{0,1\}^K$。检索时，将这些由硬件生成的编码与预先存储的数据库编码进行汉明距离匹配，以对语义相似的样本排序并进行检索。

图5展示了具有代表性的实验编码示例。DFB-SA 激光器能够可靠地将连续输入映射为离散的 0/1 脉冲输出，从而有效实现最终的光子二值化层。重要的是，所提出的量化裕度损失扩大了固定阈值 T_{out} 周围的安全裕度，抑制阈值边界处的模糊性，并减轻光子硬件部署中的噪声引起的比特翻转。

与我们此前在 [35], 中基于DFB-SA的实现相比（该实现观察到的平均错误率约为 4%），当前框架在协同设计的

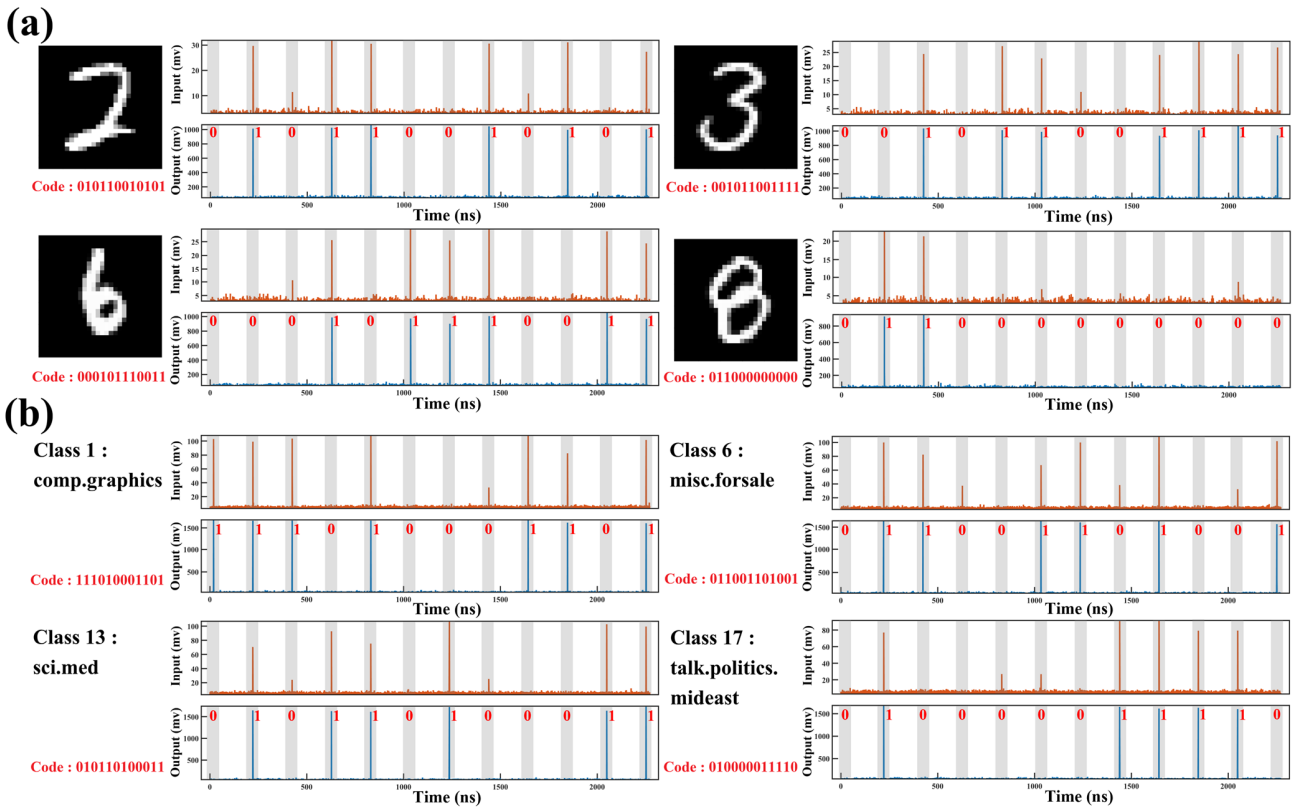

图 5. 二值化脉冲生成的软硬件实验验证。(a) MNIST 和 (b) 20NG 代表性样本的时序输入预激活（上）及其对应的二进制输出脉冲（下），用于生成最终的12位哈希码。

闭环中未观察到任何错误。具体而言，在所有测试查询（每个数据集 100 个查询）和所有生成的位（12 位编码）中，DFB-SA激光器均产生完全正确的 0/1 输出，没有发生位翻转（即，MNIST 的 100×12 位中错误数为 0，20NG 的 100×12 位中错误数也为 0）。这一结果直接证明，所提出的边界感知训练和固定阈值接口显著提升了硬件鲁棒性和可重复性。

采用环路中的物理阈值化后，端到端检索在 MNIST 上达到 $m\infty P = 0.9654$，在 20NG 上达到 $m\infty P = 0.9448$。这些结果表明，以真实的光子阈值化器件取代算法编码后，仍能保持较高的检索准确率，验证了从神经编码到硬件二值化和数据库匹配的整个流程具有一致性。

此外，我们评估了 DFB-SA编码模块的推理延迟和能效。单个 DFB-SA芯片的功耗估算为：

$$P_{\text{single}} = 1\,097 \times 26\,8 + 1\,28 \times \ldots\ldots 2\,13 \approx 32\,13$$

mW。给定自脉动频率 $f_{\text{sp}} = 5.23$ ت ن ه [36]，其电光响应延迟为：

$$\tau_{\text{DFB-SA}} = \frac{1}{f_{\text{sp}}} = 191\,20 \quad = 12 \quad 1$$

。ps。对于本研究中的 ك 位，在时分复用方案下，编码延迟为：

$$\tau_{\text{enc}} = \tau_{\text{DFB-SA}} = 12 \times 191\,20 \quad \approx 2\,294 \quad \text{ك}$$

。ps。ns。每次查询对应的能耗为：

$$\text{ق}_{\text{enc}} = P_{\text{single}} \cdot \tau_{\text{enc}} \approx 32.13\ \text{mW} \times 2.294\ \text{ns}\ 73\,70 \quad .$$

pJ/query。或者，采用并行的 ك 通道 DFB-SA阵列，可将延迟降低至 $\sim \tau_{\text{DFB-SA}}$（即 $\sim 191.2\ p$ ك），同时保持能耗处于同一量级。总体而言，这些结果表明，DFB-SA阈值处理模块能够以超低延迟生成二进制哈希码，并提供稳定的 0/1 光子脉冲输出，为大规模、高吞吐量的光子电子检索系统提供了切实可行的构建模块。

5. 结论

本研究展示了一种用于相似性检索的软硬件协同设计的光子脉冲哈希框架，该框架采用 DFB-SA 激光器进行硬件阈值化。通过采用量化裕度损失来扩大阈值间隔，我们在所有实验查询中均实现了无错误的二值输出，验证了所提策略的鲁棒性。在12位配置下，系统的编码延迟为 2.294 ns/查询，能耗为 73.70 pJ/查询，并有望在并行阵列中达到 191.2 p ك。这些结果证实了 DFB-SA 在高速低功耗硬件二值化方面的可靠性。未来的研究将把该框架扩展至更大的数据库和更高的位宽，并系统评估其在噪声和漂移等实际条件下的鲁棒性。此外，通过集成基于MZI的光计算和并行读出链路，我们希望推动高吞吐量光电子哈希检索系统的发展。最终，通过满足对高能效极致检索吞吐量的迫切需求，该架构为下一代智能系统提供了颇具前景的硬件基础——从加速大语言模型中的检索增强生成，到实现实时大规模多媒体推荐和大规模基因组同源性搜索。

资助信息。 本研究得到国家自然科学基金委员会（编号：62535015）资助。

利益披露。作者声明不存在利益冲突。

参考文献

1. P. Lewis、E. Perez、A. Piktus等，“知识密集型自然语言处理任务的检索增强生成”，Adv. Neural Inf. Process. Syst. 33，9459–9474（2020）。2. L. Zhu、C. Zheng、W. Guan等，“面向高效多媒体检索的多模态哈希：综述”，IEEE Trans. Knowl. Data Eng. 36(1)，239–260（2024）。3. K. Grauman和R. Fergus，“面向大规模图像搜索学习二进制哈希码”，载于《计算机视觉机器学习》，R. Cipolla、S. Battiato和G. M. Farinella编（Springer，2013），第411卷，第49–87页。4. A. A. Karcioglu和H. Bulut，“利用完美哈希改进面向DNA序列的hash-q精确字符串匹配算法”，Comput. Biol. Med. 131，104292（2021）。5. T. Zhang、G. Gupta、A. Desai等，“具有局部性的身份标识：一种用于基因序列搜索的理想哈希”，载于第31届ACM SIGKDD知识发现与数据挖掘大会论文集（ACM，2025），第1972–1983页。6. A. Gionis、P. I ndyk和R. Motwani，“通过哈希实现高维相似性搜索”，载于第25届超大型数据库国际会议论文集（1999），第518–529页。7. P. Indyk和R. Motwani，“近似最近邻：迈向消除维数灾难”，载于第三十届ACM计算理论年度研讨会论文集（ACM，1998），第604–613页。8. Y. Gong、S. Lazebnik、A. Gordo等，“迭代量化：一种面向大规模图像检索学习二进制编码的普罗克汝斯忒斯方法”，IEEE Trans. Pattern Anal. Mach. Intell. 35，2916–2929（2013）。9. H. Zhu、M. Long、J. Wang等，“用于高效相似性检索的深度哈希网络”，Proc. AAAI Conf. Artif. Intell. 30(1)，2415–2421（2016）。10. H. Liu、R. Wang、S. Shan等，“面向快速图像检索的深度监督哈希”，载于IEEE计算机视觉与模式识别会议论文集（2016），第2064–2072页。11. W.-J. Li、S. Wang和W.-C. Kang，“基于特征学习并带成对标签的深度监督哈希”，载于第25届国际人工智能联合会议论文集（AAAI Press，2016），第1711–1717页。12. Z. Cao、M. Long、J. Wang等，“HashNet：通过连续化进行深度哈希学习”，载于IEEE国际计算机视觉会议论文集（2017），第5608–5617页。13. R. Muralidhar、R. Borovica-Gajic和R. Buyya，“节能计算系统：从架构、抽象与建模到技术与标准”，ACM Comput. Surv. 54，1–37（2022）。14. S. R. Agrawal、C. M. Dee和A. R. Lebeck，“利用加速器实现高效的高维相似性搜索”，SIGPLAN Not. 51(8)，1–12（2016）。15. T. Fu、J. Zhang、R. Sun等，“光学神经网络：进展与挑战”，Light Sci. Appl. 13，263（2024）。16. S. Hua、E. Divita、S. Yu等，“一种具有超低延迟的集成大规模光子加速器”，《自然》640，361–367（2025）。17. X. Shao、J. Su、M. Lu等，“具有片上可集成光学平均池化的全光卷积神经网络用于图像分类”，Appl. Opt. 63，6263–6271（2024）。18. K. Liao、Y. Chen、Z. Yu等，“基于卷积神经网络的全光计算”，Sci. China Inf. Sci. 64，112205（2021）。19. G. Van Der Sande、D. Brunner和M. C. Soriano，“光子储备池计算的进展”，Nanophotonics 6，561–576（2017）。20. J. Bueno、S. Maktoobi、L. Froehly等，“大规模光子循环神经网络中的强化学习”，Optica 5，756–760（2018）。21. S. Afifi、F. Sunny、M. Nikdast等，“TRON：基于非相干硅光子学的Transformer神经网络加速”，载于五大湖VLSI研讨会论文集（ACM，2023），第15–21页。22. Y. Tian、S. Xiang、X. Guo等，“光子 Transformer 芯片：干涉即所需”，PhotoniX 6，45（2025）。23. M. Miscuglio、A. Mehrabian、Z. Hu等，“用于光子神经网络的全光非线性激活函数”，Opt. Mater. Express 8，3851–3863（2018）。24. A. Tavanaei、M. Ghodrati、S. R. Kheradpisheh等，“脉冲神经网络中的深度学习”，Neural Netw. 111，47–63（2019）。25. S. Ghosh-Dastidar和H. Adeli，“脉冲神经网络”，Int. J. Neural Syst. 19，295–308（2009）。

26. J. Feldmann、N. Youngblood、C. D. Wright等，“具有自学习能力的全光脉冲神经突触网络”，《自然》**569**，208–214（2019）。27. A. Jha、C. Huang、H.-T. Peng等，“光子脉冲神经网络与石墨烯硅基脉冲神经元”，《光波技术杂志》**40**，2901–2914（2022）。28. R. Talukder、A. Skalli、X. Porte等，“一个包含40 000个神经元的脉冲光子神经网络，采用延迟与秩序编码进行训练以利用稀疏性”，《神经形态计算与工程》**5**，034003（2025）。29. Y. Zhang、S. Xiang、C. Yu等，“在非相干光注入条件下利用脉冲式分布反馈自电注入锁定激光器实现光子神经形态模式识别”，《激光与光子学评论》**19**，2400482（2025）。30. Z. Qin、P. Zhang、F. Wu等，“Fcanet：频率通道注意力网络”，载于IEEE/CVF国际计算机视觉会议论文集（2021），第783–792页。31. S. Xiang、Y. Zhang、S. Shi等，“面向模式分类的硬件感知轻量级光子脉冲神经网络”，arXiv:2512.00419（2025）。https://doi.org/10.48550/arXiv.2512.00419。32. P. Bafna、D. Pramod和A. Vaidya，“文档聚类：词频-逆文档频率方法”，载于2016年国际电气、电子与优化技术会议（ICEEOT）（IEEE，2016），第61–66页。33. S. Deerwester、S. T. Dumais、G. W. Furnas等，“基于潜在语义分析的索引”，《美国信息科学学会杂志》**41**，391–407（1990）。34. H. Zheng、Y. Wu、L. Deng等，“借助直接训练的大型脉冲神经网络走向更深”，Proc. AAAI Conf. Artif. Intell. **35**，11062–11070（2021）。35. X. Zeng、S. Xiang、H. Zhao等，“一种面向连续控制的硬件感知光子脉冲式DDPG强化学习架构”，《激光与光子学评论》e202502481（2026）。36. S. Xiang、Y. Chen、H. Zhao等，“用于脉冲强化学习的非线性光子神经形态芯片”，arXiv:2508.06962（2025）。https://doi.org/10.48550/arXiv.2508.06962。