计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 2143−2153, 2026

SpikingReLU: 一种基于混合模型转换的高性能脉冲神经网络

王宇飞[1] 赵静月[2] 张 歌[3] 樊 斌[4] 杨智杰[2] 万 众[2] 戴玉超[4] 王 蕾[2]
[1](启元实验室 北京 100095)
[2](军事科学院国防科技创新研究院 北京 100091)
[3](北京跟踪与通信技术研究所 北京 100094)
[4](陕西省信息获取与处理重点实验室(西北工业大学) 西安 710129)
(wangyufei@qiyuanlab.com)

SpikingReLU: A High-Performance Spiking Neural Network Based on Hybrid Model Conversion

Wang Yufei[1], Zhao Jingyue[2], Zhang Ge[3], Fan Bin[4], Yang Zhijie[2], Wan Zhong[2], Dai Yuchao[4], and Wang Lei[2]
[1](*Qiyuan Lab*, *Beijing* 100095)
[2](*Institute of National Defense Science and Technology Innovation*, *Beijing* 100091)
[3](*Beijing Institute of Tracking and Telecommunication Technology*, *Beijing* 100094)
[4](*Shaanxi Key Laboratory of Information Acquisition and Processing* (*Northwestern Polytechnical University*), *Xi'an* 710129)

Abstract Brain-inspired spiking neural networks (SNNs) have shown considerable promise due to their strong spatiotemporal encoding capabilities and event-driven computation. Current mainstream SNN approaches can be divided into two categories: converting pre-trained artificial neural networks (ANNs) into SNNs (ANN conversion methods) and directly training SNNs. However, ANN conversion methods require large time steps to reduce conversion errors, while directly trained SNNs suffer from poor representational capacity. To address these challenges, we propose a high-performance spiking neural network based on hybrid model conversion. The proposed method decouples model training and inference. During the training phase, a portion of the spiking neurons in the SNNs are strategically replaced with ReLU activation functions to enhance the feature learning ability, thereby constructing a high-performance hybrid model that integrates both ANN and SNN components. Since the ANN part of the hybrid model cannot perform event-driven computation, we further introduce a reparameterization technique to convert the entire inference process of the hybrid model into event-driven computation without performance loss. Therefore, the proposed method combines the advantages of high performance and event-driven computation. Experimental results show that the proposed method achieves classification accuracies of 97.31% and 83.34% on CIFAR-10 and CIFAR-100, respectively, 70.89% on ImageNet, and 82.71% on the neuromorphic dataset CIFAR10-DVS, outperforming state-of-the-art methods.

Key words brain-inspired computing; spiking neural networks; event-driven computing; hybrid model; training-inference decoupling method

摘 要 受大脑启发的脉冲神经网络(spiking neural networks, SNNs)由于其强大的时空编码能力和事件驱动计算等特性展现出巨大的发展潜力。目前,主流的SNNs方法分为将训练完备的人工神经网络(artificial neural networks, ANNs)转换为SNNs以及直接训练SNNs这2类。然而,现有的ANNs转换方

收稿日期: 2025-09-30; **修回日期**: 2026-03-30
基金项目: 国家自然科学基金项目(62372461, 62032001, 62203457, 62406335)
This work was supported by the National Natural Science Foundation of China (62372461, 62032001, 62203457, 62406335).
通信作者: 王蕾(leiwang@nudt.edu.cn)

法需要较长的时间步减少转换误差,直接训练的 SNNs 则受限于网络表征能力的不足。针对上述挑战,提出一种基于混合模型转换的高性能脉冲神经网络。所提方法将模型训练与推理解耦:在训练阶段,策略性地将 SNNs 中部分脉冲神经元替换为 ReLU 激活函数,增强网络的特征学习能力,从而构建一个同时包含 ANNs 与 SNNs 的高性能混合模型;由于混合模型中的 ANNs 部分无法进行事件驱动计算,进一步引入重参数化技巧,在不损失性能的前提下,将混合模型的推理过程全部转换为事件驱动计算,从而使所提方法兼具高性能和事件驱动计算的优势。实验结果表明:所提方法在 CIFAR-10 和 CIFAR-100 数据集上的分类准确率分别为 97.31% 和 83.34%,在 ImageNet 数据集上的分类准确率为 70.89%,在神经静态数据集 CIFAR10-DVS 的分类准确率为 82.71%,均优于现有国内外最先进的方法。

关键词 类脑计算;脉冲神经网络;事件驱动计算;混合模型;训练-推理解耦方法

中图法分类号 TP389.1; TP391.4

DOI: 10.7544/issn1000-1239.202550722 **CSTR:** 32373.14.issn1000-1239.202550722

类脑计算旨在借鉴生物神经系统的结构特征与信息处理机制,突破传统冯·诺依曼架构在能效方面的瓶颈,构建新一代高效智能计算体系[1]。脉冲神经网络(spiking neural networks, SNNs)凭借其高度仿生的神经元模型和事件驱动计算机制,受到国内外研究者的广泛关注,是类脑计算领域的重要研究方向[2]。相对于人工神经网络(artificial neural networks, ANNs)中的非线性激活层[3], SNNs 使用脉冲神经元作为激活层,赋予了 SNNs 强大的时空编码能力和事件驱动计算等特性。

主流的 SNNs 研究大致分为 ANNs 转换方法与直接训练 SNNs 这 2 类。然而,现有的 ANNs 转换方法依赖数百个时间步来减小转换误差,导致推理效率低下;直接训练的 SNNs 受限于现有神经元较弱的表征能力,难以突破性能瓶颈。为提升 SNNs 的性能,国内外研究学者提出学习神经元的关键参数,如脉冲阈值[4]和膜时间常数[5]。Ternary spike[6]和 I-LIF[7]提出将 2 值神经元改进为 3 值和整数值。受 ANNs 启发,基于 SNNs 的残差学习网络[8-9]和注意力机制[10]也相继被提出以缓解网络深度退化问题并提高性能。此外,一些研究者提出施加人为设计的正则化[11]或使用性能较强的 ANNs 模型进行知识蒸馏[12]。

尽管现有研究已取得显著进展,但存在一个关键问题尚未得到充分重视。如图 1(a)所示,主流的 SNNs 网络主干通常叠加多个“卷积-批量归一化-脉冲神经元”(convolution batch-normal spiking-neuron, CBS)模块,然而,本文实验发现,多个堆叠的 CBS 模块并不能有效提高 SNNs 性能。具体而言,CBS 模块中的卷积层应尽可能从脉冲输入中提取高级特征。但该模块中的单层卷积却只能执行简单的线性操作,学习能力有限,无法有效地从脉冲输入中提取复杂特征。同时,CBS 模块的神经元又将卷积层的输出重新量化为脉冲特征,导致后续的 CBS 模块会遇到相同的限制。

为解决上述问题,本文提出一种基于混合模型转换的脉冲神经网络——SpikingReLU,使 SNNs 具备高性能和事件驱动计算的双重优势。具体而言,SpikingReLU 策略性地将多个连续 CBS 模块序列中除最后一个神经元之外的所有神经元替换为 ReLU 激活层,其中最后一个神经元确保模块的输出仍为脉冲信号。多个结合非线性 ReLU 激活层的卷积层赋予了网络强大的特征提取能力,但也破坏 SNNs 事件驱动计算的优势。如图 1(b)所示,针对该问题,本文进一步提出一种训练-推理解耦策略,首先训练得到一个同时包含 ANNs 和 SNNs 的高性能混合网络模型;在推理阶段,采用重参数化技巧重构模型计算过程,将其完全转换为事件驱动计算。

综上所述,本文的主要贡献如下:

1)提出了一个同时包含 ANNs 和 SNNs 的混合模型,策略性地将现有 SNNs 中的部分神经元替换为非线性 ReLU 激活层,显著提高了网络的特征表征能力;

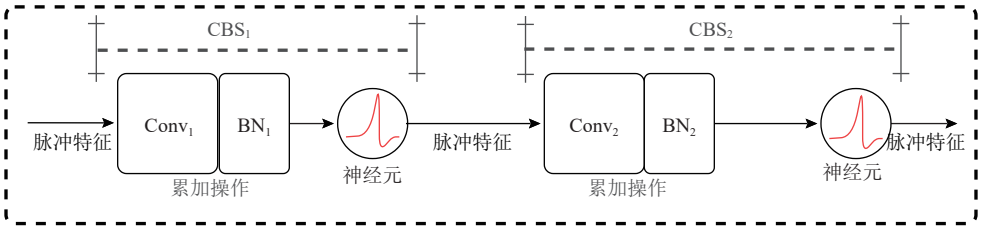
2)提出了一种无损的混合模型转换方法,在模型推理阶段将混合模型完全转换为事件驱动计算,从而使混合模型兼具高性能和事件驱动计算的优势;

3)在多个主流的静态数据集与神经形态数据集上的实验结果表明,本文方法的性能优于现有国内外最先进方法。

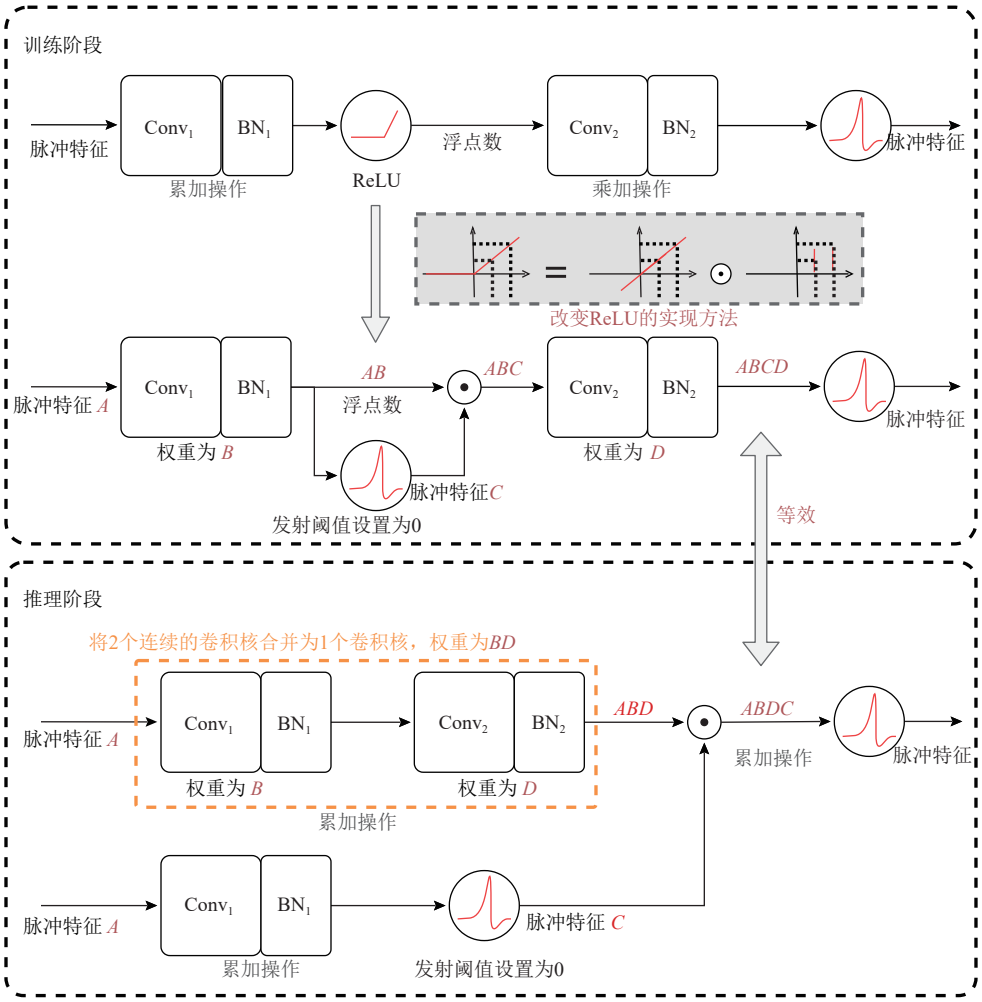
1 相关工作

1.1 ANNs 转换方法

ANNs 转换方法通过将训练完备的 ANNs 转换



(a) SNNs中传统的CBS脉冲特征提取模块



(b) 本文提出的SpikingReLU脉冲特征提取模块

Fig. 1 The proposed SpikingReLU module and traditional CBS module

图 1 本文提出的 SpikingReLU 模型与传统 CBS 模块

为 SNNs, 从而快速且有效地获取 SNNs 模型。该领域开创性的工作[13]验证了 ReLU 激活函数与脉冲神经元发放率之间的等价性, 并将学习到的 ANNs 模型权重转移到 SNNs 模型中。文献 [14−15] 首先移除 ANNs 网络的局部响应归一化层, 将最大值池化改为平均池化, 之后训练使用软化泄漏积分发放(leaky integrate-and-fire, LIF)响应函数作为非线性层的网络, 并在训练过程中添加噪声增强网络的鲁棒性, 最后使用 LIF 脉冲模型替换软化的 LIF 率模型, 从而将 ANNs 网络转换为 SNNs。之后, 文献 [16] 提出在 ANNs 训练阶段使用零偏差的修正线性单元, 在推理阶段使用一种新的权重归一化方法来帮助调节神经元发放率, 从而减少卷积网络和全连接深度网络在转换过程中的性能损失。为进一步减少转换误差, SATP[17]提出了一种自驱动自适应阈值可塑性机制, 其中神经元使用无监督学习规则并根据其自身的状态信息自主调整发放阈值。为减少转换之后 SNNs 的时间步长, 文献 [18] 对 ANNs 到 SNNs 的转换过程进行了理论分析, 表明缩放阈值发挥着与权重正则化类似的作用, 基于此理论结果, 提出了优化初始膜电位减

少每层的转换损失,将时间步缩小为 32。

1.2 直接训练 SNNs 方法

由于 SNNs 中的神经元使用阶跃函数判定是否发射脉冲,阶跃函数在脉冲发射点处的导数为无穷大,而其他地方的导数为 0,这导致 SNNs 无法直接使用反向传播对权重进行更新。针对该问题,现有方法在反向传播时使用一个光滑可微的替代函数来近似不可微的阶跃函数,使 SNNs 可以直接训练。直接训练的 SNNs 模型主要从神经元、网络结构和训练技术 3 个层面提升脉冲神经网络的性能[2]。在神经元层面,研究人员提出了 PLIF[5], GLIF[19] 等多种改进神经元模型。在网络结构层面,受 ANNs 中有效网络结构的启发,SNNs 领域出现了脉冲版本的残差学习网络[8-9]、注意力机制[10] 和 Transformer 架构[20],显著提升了网络的性能。在训练技术层面,研究者通过引入辅助损失函数作为正则化项[11],或在学习阶段使用更强大的 ANNs 教师模型来引导 SNNs 模型进行知识蒸馏[12]。此外,Real spike[21] 和 ILIF[7] 等方法提出了训练-推理解耦的策略,在保持事件驱动优势的同时,有效提升了直接训练 SNNs 的性能。

2 预备知识

2.1 LIF 神经元

神经元是脉冲神经网络的核心计算单元,其主要功能是整合时空特征并将其量化为脉冲。其中,LIF 神经元因其可以较好地平衡生物合理性和计算复杂度,是目前最常用的神经元模型。如式(1)所示,LIF 神经元的神经学动力过程包含 3 个步骤:充电(charging)、发放(firing)和重置(resetting)。具体而言,在时间步 t,来自前一时刻的膜电位 $\boldsymbol{V}[t-1]$ 与当前的空间输入信息 $\boldsymbol{X}[t]$ 被整合到当前膜电位 $\boldsymbol{H}[t]$。脉冲输出 $\boldsymbol{B}[t]$ 由阶跃函数 $\Theta(\cdot)$ 决定:当膜电位 $\boldsymbol{H}[t]$ 超过发放阈值 $\boldsymbol{V}_{\text{threshold}}$ 时,LIF 神经元产生一个脉冲($\boldsymbol{B}[t]=1$),否则保持静息状态($\boldsymbol{B}[t]=0$)。发放脉冲会消耗神经元累积的电荷,导致膜电位瞬时下降,即膜电位重置过程。以常用的软重置为例,若神经元发放脉冲,则从膜电位中减去阈值电压 $\boldsymbol{V}_{\text{threshold}}$。为简化表达,本文将 LIF 神经元的充电过程记作函数 $Charging(\cdot)$,发放过程记作函数 $Firing(\cdot)$,重置过程记作函数 $Resetting(\cdot)$。

$$LIF=\begin{cases} Charging:\boldsymbol{H}[t]=\boldsymbol{V}[t-1]+\boldsymbol{X}[t], \\ Firing:\boldsymbol{B}[t]=\Theta(\boldsymbol{H}[t]-\boldsymbol{V}_{\text{threshold}}), \\ Resetting:\boldsymbol{V}[t]=\boldsymbol{H}[t]-\boldsymbol{V}_{\text{threshold}}\boldsymbol{B}[t]。 \end{cases} \tag{1}$$

2.2 CBS 脉冲特征提取模块

基于目前成熟的 ANNs, SNNs 致力于用 LIF 神经元替代 ANNs 中的非线性激活层。因此,与 ANNs 中常见的“卷积-批归一化-非线性激活”(convolution batch-normal non-linear, CBN)特征提取模块不同,SNNs 通常由多个堆叠的 CBS 模块构成。

对于一个 CBS 特征提取模块,假设输入特征图为 $\boldsymbol{B}^0$,其特征图的高度和宽度分别为 H 和 W,通道数为 C。CBS 模块中的卷积层的核为 $\boldsymbol{K}^1$,卷积核的维度为 $KKCC$,其中 K 为卷积核尺寸大小,输入与输出特征图的通道数都为 C,该卷积层的输出特征图 $\boldsymbol{F}^1$ 中每个元素的计算方式为:

$$\boldsymbol{F}^1_{k,l,n}=\sum_{i,j,m}\boldsymbol{K}^1_{i,j,m,n}\cdot\boldsymbol{B}^0_{k+i-1,l+j-1,m}, \tag{2}$$

其中输入特征图 $\boldsymbol{B}^0$ 通常是由前一个 CBS 模块生成或为输入信号经过脉冲编码得到的二进制脉冲特征。在神经形态硬件上,CBS 模块中的卷积运算可以通过累加(accumulation, AC)操作实现,相比于传统卷积的乘加(multiply-accumulation, MAC)操作更加节能[22-24],这是 SNNs 实现高能效的一个关键因素。为了确保训练稳定性,卷积层之后会使用一个批归一化(batch normalization, BN)层。但由于 BN 层在推理过程中可以与前一层卷积融合为一个带偏置的卷积层,本文不对 BN 层进行单独描述。最后,卷积层的输出被送入 LIF 神经元,将浮点特征重新量化为二进制脉冲特征,其过程如式(3)所示。

$$\boldsymbol{B}^1=Firing(\boldsymbol{F}^1)。 \tag{3}$$

3 混合模型转换方法

3.1 SpikingReLU 脉冲特征提取模块

传统的 CBS 脉冲特征提取模块存在 2 个缺陷:首先,卷积层的输出特征量化为脉冲信号的过程会产生显著的量化误差,阻碍特征信息的传播;其次,单层卷积是一个学习能力有限的线性操作,无法有效地从脉冲输入信号中提取到高级特征。为解决上述问题,本文提出一种混合 ANNs 和 SNNs 特性的脉冲特征提取模块—SpikingReLU,提出的模型可作为现有 CBS 脉冲特征提取模块的改进。以改进 2 个连续的 CBS 模块为例,SpikingReLU 将 CBS 模块序列中除最后一个神经元之外的所有神经元替换为非线性 ReLU 激活层,其中最后一个神经元确保 SpikingReLU 模块的输出仍为脉冲信号。如式(4)所示,SpikingReLU

将第 1 个 CBS 模型中卷积层的输出特征 $\boldsymbol{F}^1$ 输入到 ReLU 激活层中，得到激活值 $\boldsymbol{G}^1$。

$$\boldsymbol{G}^1 = ReLU(\boldsymbol{F}^1)。 \tag{4}$$

之后，将其输入到第 2 个 CBS 模块的卷积层中，假定该卷积层的核为 $\boldsymbol{K}^2$，输出特征记为 $\boldsymbol{F}^2$，计算过程如式(5)所示。

$$\boldsymbol{F}^2_{k,l,n} = \sum_{i,j,m} \boldsymbol{K}^2_{i,j,m,n} \cdot \boldsymbol{G}^1_{k+i-1,l+j-1,m}, \tag{5}$$

其中 ReLU 激活层的输出特征 $\boldsymbol{G}^1$ 由浮点数组成，因此后续的 CBS 模块中的卷积层均无法以事件驱动的方式计算。

3.2 训练-推理解耦策略

为解决上述问题，本文提出一种训练-推理解耦策略，如图 2 所示。在模型训练阶段，SpikingReLU 按照式(2)(4)(5)顺序执行。在模型推理阶段，SpikingReLU 采用重参数化技术将混合的 ANNs 和 SNN 模型转换为完全事件驱动的计算方式，从而使模型兼具高性能和事件驱动计算的优势。

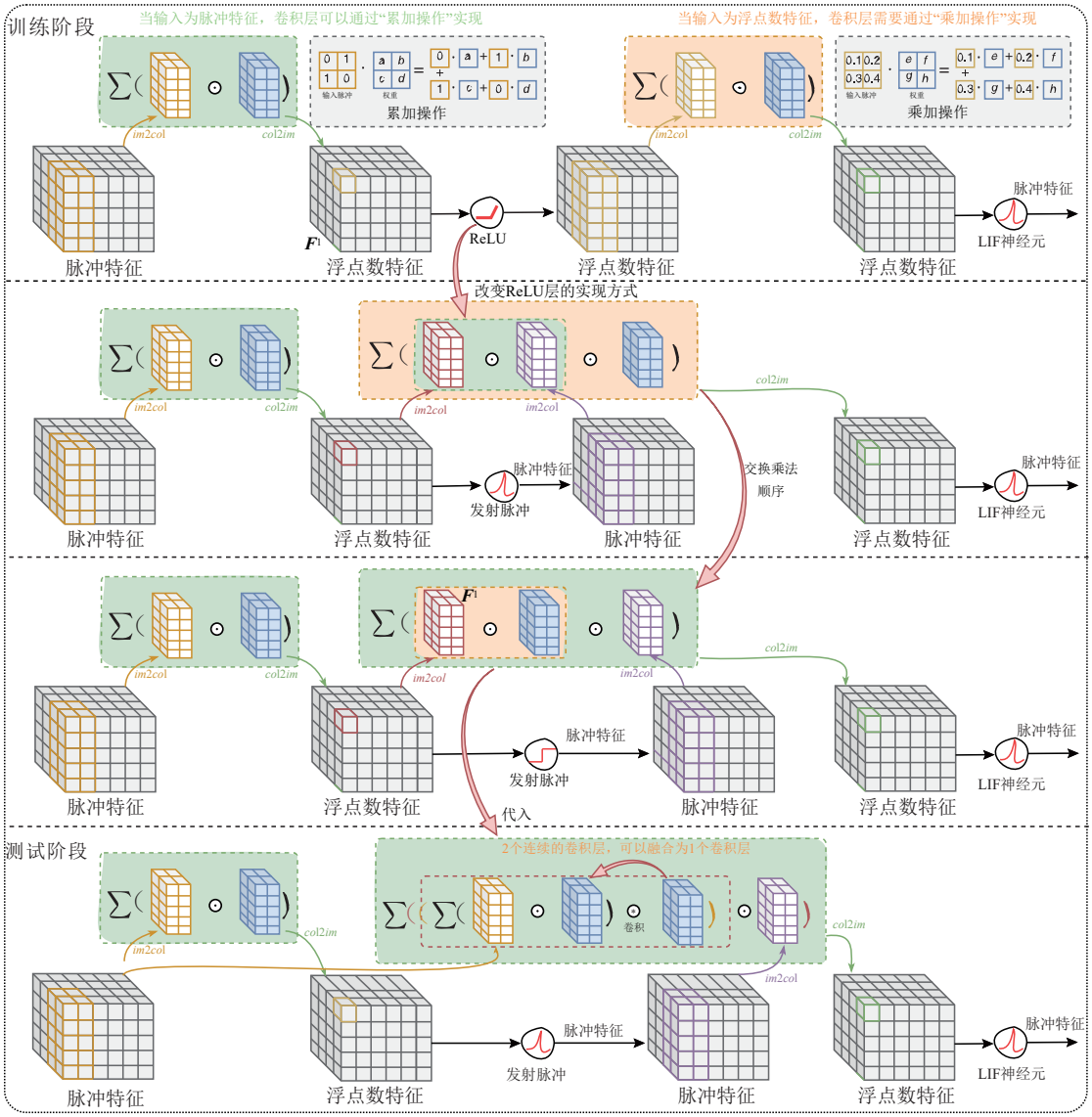


Fig. 2 The schematic diagram of hybrid model conversion based on the training-inference decoupling strategy

图 2 基于训练-推理解耦策略的混合模型转换示意图

具体而言，在模型推理阶段，首先改变非线性 ReLU 激活层原本的计算方式。对于输入特征 $\boldsymbol{F}^1$，ReLU 激活层的功能是将其所有负值转换为零，同时保持正值不变。如式(6)所示，$\boldsymbol{F}^1$ 的正值掩码 $\boldsymbol{M}^1$ 可以通过将发射阈值设为 0 的发射操作获得。

$$\boldsymbol{M}^1 = Firing_{\text{threshold}=0}(\boldsymbol{F}^1)。 \tag{6}$$

因此，与 ReLU 激活层原本的计算方式不同，其激活特征 $\boldsymbol{G}^1$ 可以通过式(7)计算得到。

$$\boldsymbol{G}^1 = \boldsymbol{F}^1 \odot \boldsymbol{M}^1 , \tag{7}$$

其中⊙操作表示逐元素相乘。

为了在模型推理阶段，将 SpikingReLU 转换为完全的事件驱动计算，首先使用式(7)中改进后的 ReLU 层计算方法替换其原本计算方法，然后将式(7)代入式(5)中，得到第 2 个 CBS 模块卷积层的输出特征 $\boldsymbol{F}^2$，具体计算过程如式(8)所示。

$$\boldsymbol{F}^2_{k,l,n} = \sum_{i,j,m} \boldsymbol{K}^2_{i,j,m,n} \cdot \left(\boldsymbol{F}^1_{k+i-1,l+j-1,m} \cdot \boldsymbol{M}^1_{k+i-1,l+j-1,m} \right) 。 \tag{8}$$

根据乘法交换律，交换式(8)中乘法的顺序，得到式(9)。

$$\boldsymbol{F}^2_{k,l,n} = \sum_{i,j,m} \left(\boldsymbol{K}^2_{i,j,m,n} \cdot \boldsymbol{F}^1_{k+i-1,l+j-1,m} \right) \cdot \boldsymbol{M}^1_{k+i-1,l+j-1,m} , \tag{9}$$

其中 $\boldsymbol{F}^1_{k+i-1,l+j-1,m}$ 是通过式(2)计算得到。因此，将式(2)代入式(9)中可得式(10)。

$$\boldsymbol{F}^2_{k,l,n} = \sum_{i,j,m} \left(\boldsymbol{K}^2_{i,j,m,n} \left(\sum_{i',j',m'} \boldsymbol{K}^1_{i',j',m',n'} \boldsymbol{B}^0_{k+i-1+i'-1,l+j-1+j'-1,m'} \right) \right) \times \boldsymbol{M}^1_{k+i-1,l+j-1,m} 。 \tag{10}$$

在式(10)中，脉冲输入信号 $\boldsymbol{B}^0$ 连续与卷积核 $\boldsymbol{K}^1$ 和卷积核 $\boldsymbol{K}^2$ 进行卷积运算，这 2 个连续的卷积层可以通过合并重构为 1 个卷积核 $\tilde{\boldsymbol{K}}$，因此计算过程可整理为如下所示：

$$\boldsymbol{F}^2_{k,l,n} = \sum_{i,j,m,i',j',m'} \left(\tilde{\boldsymbol{K}}_{i',j',m',n} \cdot \boldsymbol{B}^0_{k+i'-1,l+j'-1,m'} \right) \cdot \boldsymbol{M}^1_{k+i-1,l+j-1,m} 。 \tag{11}$$

在式(11)中，特征 $\boldsymbol{B}^0$ 和 $\boldsymbol{M}^1$ 都由二进制脉冲信号组成，因此，SpikingReLU 在推理阶段，所有运算都可以采用事件驱动的计算方式执行。

3.3 用法与扩展

3.1 节和 3.2 节以 2 个连续的 CBS 脉冲特征提取模块为例，介绍了 SpikingReLU 改进方法。值得一提的是，SpikingReLU 对于多个连续的 CBS 模块同样有效。具体而言，在模型训练阶段，SpikingReLU 将 CBS 模块序列中除最后一个之外的神经元替换为 ReLU 激活层，并在推理阶段采用重参数化技术将 SpikingReLU 转换为纯事件驱动计算。表 1 的实验结果表明：当模块数量增加时，SpikingReLU 模块比传统 CBS 模块获得了更显著的性能提升，能够更有效地缓解网络深度退化问题。此外，实验结果还表明 SpikingReLU 模块中还可采用一些改进的 LIF 神经元，如 3 值脉冲神经元和 I-LIF 神经元，协同提升 SNNs 的性能。

本文提出的 SpikingReLU 模块作为对传统 CBS 模块的改进，可用于替换现有 SNNs 网络中 2 个或多个连续的 CBS 模块。例如，SpikingReLU 模块可在输入信号脉冲编码阶段使用，也可用于基于 SNNs 的残差网络中学习残差特征图。

Table 1 Ablation Studies for Accuracy Results on CIFAR-10 Dataset

表 1 在CIFAR-10数据集上的消融实验准确率结果 %

模型	神经元	模块的数目		
		1	2	3
CBS 模块	2 值	95.82	96.19	96.25
	Real spike[21]	95.85	96.31	96.34
SpikingReLU	2 值	95.82	97.05	97.31
	3 值			97.47
	4 值			97.46

3.4 讨 论

本节对比了提出的 SpikingReLU 与现有的相关工作。早期研究[14-15]提出一种替换 LIF 神经元的训练方法，首先修改 ANNs 的网络结构，使用软化 LIF 神经元模型并配合噪声训练得到静态网络，之后将软化 LIF 神经元模型替换为真正的 LIF 神经元，得到性能相近的 SNNs。与之相比，SpikingReLU 不需要对 ANNs 网络进行修改，在训练阶段策略性地使用 ReLU 替换 LIF 神经元，在推理阶段使用重参数技巧实现脉冲驱动计算。Real spike[21]首次提出了一种训练-推理解耦的脉冲神经网络。在训练阶段，Real spike 为 2 值脉冲预测一个系数图以获得实数值脉冲；在推理阶段，该系数图可以通过使用重参数化技术被吸收到后续的卷积核中。Real spike 中的系数图在训练阶段中通过学习优化，在推理阶段保持不变。相比于提出的 SpikingReLU，特征图 $\boldsymbol{F}^1$ 工作原理类似于 Real spike 的系数图。但是，$\boldsymbol{F}^1$ 不同于 Real spike 的可学习的超参数，而是由前一层卷积预测得到。因此，SpikingReLU 的系数图 $\boldsymbol{F}^1$ 可根据输入样本进行自适应调整，从而进一步增强 SNNs 的表征能力。

4 实验结果与分析

4.1 数据集和评价指标

为了全面验证 SpikingReLU 的优越性，本文在 RGB 图像和事件流 2 种不同数据进行实验。在 RGB 图像分类实验中，选择了当前主流的图像分类数据集，其中包括 CIFAR-10 和 CIFAR-100 以及 ImageNet

数据集。在对事件流实验中，选择了常用的 CIFAR10-DVS 数据集。遵循通用的评估方法，将计算分类准确率作为分类方法的主要性能指标。

4.2 方法实现细节

本文模型使用 Pytorch 实现，并在 GeForce RTX 3090 GPU 上进行实验。本文的所有实验均使用 SGD 优化器，动量设置为 0.9，学习率遵循余弦退火衰减策略。对于 RGB 图像分类任务，批大小（batch size）设置为 128，SpikingReLU 在 CIFAR-10 和 CIFAR-100 数据集上训练 400 轮，在 ImageNet 数据集上训练 100 轮。此外，本文还采用一些常见策略来提高方法的性能，包括自动数据增强和测试时增强。对于事件数据，使用 SpikingJelly 软件库加载事件数据，批大小设置为 12，在 CIFAR10-DVS 数据集上训练 200 轮。

4.3 SpikingReLU 与传统 CBS 模块的对比

表 1 中对比了基于 SNNs 的残差学习网络模块以及传统 CBS 模块的性能。实验结果表明，当残差学习模块仅使用单个 CBS 模块学习残差图时，分类准确率为 95.82%，当将 CBS 模块数量增加到 2 时，准确率为 96.19%，性能提升相对较小。表 1 同时表明，由于实数值 LIF 神经元比 2 值 LIF 具有更强的特征表达能力，使用 Real spike 可以提高方法的准确率。值得注意地是，使用本文所提出的 SpikingReLU 改进 2 个连续的 CBS 模块，SNNs 的性能会显著增强，这是由于 SpikingReLU 采用多个卷积层和非线性层可以有效地从脉冲输入中学习高级特征表示。改进之后方法的准确率为 97.05%，比使用 CBS 模块的方法高 0.86 个百分点，比使用 CBS 与 Real spike 结合的方法高 0.75 个百分点。同时，随着模块数量的增加，SpikingReLU 的性能提升幅度更加显著，进一步证明了其有效性。此外，图 3 中可视化了 CBS 模块和 SpikingReLU 模块在不同时间步的中间特征的同一个通道特征。从中观察到，CBS 模块的特征是离散的，仅由 0 和 1 组成，从而限制了 CBS 模块之间信息的传递，对比而言，SpikingReLU 模块中的特征更加丰富。

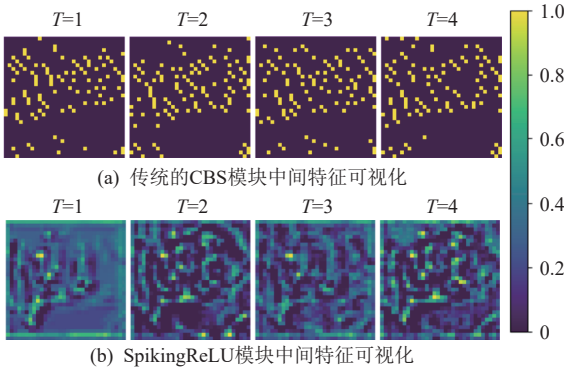


Fig. 3 Feature visualization diagrams at different time steps

图 3 不同时间步下的特征可视化图

4.4 不同时间步对比

对于 SNNs，更大的时间步意味着更强的时空特征提取能力。遵循现有工作，本文将时间步设置为 1，2，4，在 CIFAR-10 和 CIFAR-100 数据集上的实验结果如表 2 所示。表 2 表明，随着时间步的增加，SpikingReLU 的准确率持续提高。值得注意的是，SpikingReLU 在相同时间步下的性能均优于现有国内外最先进的方法，这充分验证了 SpikingReLU 在相同时间步设置下能够更好地利用时空特征。

4.5 不同神经元对比

近年来，国内外的研究学者提出了一系列改进的 LIF 神经元来提升 SNNs 的性能，其中最具代表性的有 3 值脉冲神经元和 I-LIF 神经元等。为验证 SpikingReLU 与改进的 LIF 神经元的兼容性，本节验证了 SpikingReLU 中采用 3 值和 4 值的 I-LIF 神经元在 CIFAR-10 数据集的性能。表 1 中的实验结果表明，当神经元从 2 值变为 3 值时，准确率从 97.31% 提高到 97.47%，表明 SpikingReLU 可以与现有的改进脉冲神经元协同作用，共同提升 SNNs 的网络表达能力。当 SpikingReLU 采用 4 值神经元时，方法的性能提升有限，这表明改进神经元带来的提升正在接近饱和。

4.6 与国内外先进方法对比

为验证 SpikingReLU 在性能上的优越性，本文在 CIFAR-10，CIFAR-100，ImageNet 这 3 个主流的图像分类数据集上，对比 SpikingReLU 与国内外先进方法的性能。本文设计的图像分类网络采用 ResNet 网络架构，并在残差学习分支中使用 SpikingReLU 模块。在 CIFAR-10 和 CIFAR-100 数据集上，SpikingReLU 的模型大小与 ResNet-18 相当，在 ImageNet 数据集上，SpikingReLU 的模型大小与 ResNet-34 相当。本文在 CIFAR-10 和 CIFAR-100 数据集上将时间步设置为 1，2，4，在 ImageNet 数据集，时间步设置为 4。为了表述简洁，本文只报告了对比方法的最优结果：表 2 表明，SpikingReLU 在 CIFAR-10 数据集上达到了 97.31% 的准确率，超越了所有对比方法，在 CIFAR-100 数据集上达到了 83.34% 的最高准确率，同样排名第 1。值得注意地是，SpikingReLU 使用所有时间步的准确率均优于现有国内外最优的方法。同时，表 3 对比了 TET[35]，RecDis-SNN[36]，GLIF[19]，Real spike[21]，RMP-loss[37]，MPBN[38]，Ternary Spike[6]，AGMM[33]，STAA-

Table 2 Performance Results Compared with State-of-the-Art Methods on CIFAR-10 and CIFAR-100 Datasets

表 2 在 CIFAR-10 和 CIFAR-100 数据集上与最先进方法的性能对比结果

类型	方法	时间步	准确率/%（CIFAR-10）	准确率/%（CIFAR-100）	发表渠道
ANNs 转换方法	SpikeNorm[25]	2 500	91.55		Front. Neurosci
	RMP[26]	2 048		67.82	CVPR
	Cbp-Qsnn[27]	32	91.51	66.63	JETCAS
直接训练 SNNs	Real spike[21]	10	95.78	71.24	ECCV
	DSR[28]	20	95.40	78.50	CVPR
	SpikeFormer[29]	4	95.51	78.21	ICLR
	SD Transformer[30]	4	95.60	78.40	NeurIPS
	Ternary Spike[6]	2	94.48	73.41	AAAI
		4	94.96	74.02	
	SWformer[31]	6	96.30	79.60	ECCV
	SNN-Vit[32]		96.10	80.10	ICLR
	AGMM[33]	2	96.13	80.25	AAAI
		4	96.33	80.70	
	Spiking Transformer[20]	4	96.42	79.90	CVPR
	STAA-SNN[34]	1	96.75	79.37	CVPR
		2	96.85	80.57	
		4	97.14	82.05	
混合模型转换	SpikingReLU（本文）	1	96.75	82.42	
		2	97.15	83.03	
		4	**97.31**	**83.34**	

注：黑体数值表示最优值。

SNN[34], FATA-SNN[39] 算法在 ImageNet 数据集的实验结果。表 3 表明，SpikingReLU 在 ImageNet 数据集上也达到了 70.89% 的最高准确率，进一步证明了 SpikingReLU 的有效性。

构建可以部署在神经形态硬件上处理来自动态视觉传感器（dynamic vision sensor, DVS）数据的脉冲神经网络是一个具有巨大应用潜力的研究方向。因此，本文进一步地在 CIFAR10-DVS 数据集上进行实验。表 4 中对比了 IM-loss[40], SpikeFormer[29], MPBN[38], Ternary spike[6], AGMM[33], STAA-SNN[34], FATA-SNN[39] 算法在 CIFAR10-DVS 数据集的实验结果。如表 4 所示，与现有国内外最优方法相比，Spiking-

Table 3 Comparative Results for Classification Accuracy on ImageNet Dataset

表 3 在 ImageNet 数据集上的分类准确率对比结果

方法	时间步	准确率/%
TET[35]	6	64.79
RecDis-SNN[36]	6	67.33
GLIF[19]	4	67.52
Real spike[21]	4	67.69
RMP-loss[37]	4	65.17
MPBN[38]	4	64.71
Ternary Spike[6]	4	70.74
AGMM[33]	4	64.67
STAA-SNN[34]	4	70.04
FATA-SNN[39]	4	70.23
SpikingReLU（本文）	4	70.89

Table 4 Comparative Results for Classification Accuracy on CIFAR10-DVS Dataset

表 4 在 CIFAR10-DVS 数据集上的分类准确率对比结果

方法	时间步	准确率/%
IM-loss[40]	10	72.60
SpikeFormer[29]	16	80.90
MPBN[38]	10	78.70
Ternary spike[6]	10	79.80
AGMM[33]	4	82.40
STAA-SNN[34]	16	82.10
FATA-SNN[39]	16	82.70
SpikingReLU（本文）	16	82.71

ReLU 的准确率最高，这表明 SpikingReLU 可以更加有效地处理事件流数据。

4.7 功耗对比

SNNs 使用二进制脉冲信号传递信息是其低功耗优势的主要来源。具体而言，由于二进制脉冲信号由 0 和 1 组成，具有稀疏性；其次，由于 SNNs 可以通过更加节能的累加操作实现。例如，在 45 nm 工艺下，累加操作能效为 0.9 pJ，传统的乘加操作能耗为 4.6 pJ。式(12)(13)为 ANNs 和 SNN 网络的功耗计算方法。

$$E_{\mathrm{ANN}} = N^2 \cdot C_{\mathrm{in}} \cdot C_{\mathrm{out}} \cdot K^2 \cdot E_{\mathrm{MAC}}\,, \tag{12}$$

$$E_{\mathrm{SNN}} = T \cdot fr \cdot N^2 \cdot C_{\mathrm{in}} \cdot C_{\mathrm{out}} \cdot K^2 \cdot E_{\mathrm{AC}}\,, \tag{13}$$

其中 N 为输出特征图的尺寸，C_{in} 为输入特征的通道数，C_{out} 为输出特征的通道数，K 为卷积核的尺寸，T 为 SNNs 的时间步大小，fr 为 SNNs 中脉冲信号的发射率。根据式(12)(13)，分别计算了 ANN(不采用脉冲驱动计算)以及 SNN(采用脉冲驱动计算)的功耗。如表 5 所示，采用 SNN 可以有效地降低功耗。

Table 5 Power Comparison on CIFAR10 Dataset

表 5 在 CIFAR10 数据集上的功耗对比 mJ

方法	功耗
ANN	239.2
SNN（本文）	87.3

5 结 论

本文提出了一种基于混合模型转换的高性能脉冲神经网络——SpikingReLU，该方法使用训练-推理解耦的策略，使 SNNs 具备高性能和事件驱动计算的优势。在训练期间，SpikingReLU 将脉冲神经网络中部分神经元替换为 ReLU 激活层，多个结合了非线性激活层的卷积层可以显著增强脉冲神经网络的表征能力，但也破坏了其事件驱动计算的优势。针对这一问题，本文进一步提出使用重参数化技术，将包含 ANNs 和 SNNs 的高性能混合网络模型完全转换事件驱动计算。在多个主流的静态和神经形态数据集上的实验结果均充分验证了本文方法的有效性和先进性。

作者贡献声明：王宇飞负责研究方案实施、实验开展与结果分析，并完成论文撰写；赵静月参与算法实验验证，负责实验结果整理及论文内容修订；张歌负责研究思路把关、理论分析指导及论文审阅；樊斌参与理论框架论证，并对论文内容和技术表述进行修改完善；杨智杰负责实验方案设计、实验流程优化及相关验证工作；万众负责论文审校、内容修改及文字规范化；戴玉超为研究提供理论指导，并参与论文结构调整与学术内容修订；王蕾负责相关理论问题指导，并对论文内容进行审阅和完善。

参 考 文 献

[1] Huang Tiejun, Yu Zhaofei, Liu Yijun. A survey of brain-inspired machine thought and architecture[J]. Journal of Computer Research and Development, 2019, 56(6): 1135−1148 (in Chinese)
(黄铁军, 余肇飞, 刘怡俊. 类脑机的思想与体系结构综述[J]. 计算机研究与发展, 2019, 56(6): 1135−1148)

[2] Guo Yufei, Huang Xuhui, Ma Zhe. Direct learning-based deep spiking neural networks: A review[J]. Frontiers in Neuroscience, 2023, 17(1): 120−795

[3] Wu Yuchen, Feng Junwen. Development and application of artificial neural network[J]. Wireless Personal Communications, 2018, 102(2): 1645−1656

[4] Bellec G, Salaj D, Subramoney A, et al. Long short-term memory and learning-to-learn in networks of spiking neurons[C]//Proc of the 32nd Int Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates Inc, 2018: 787−797

[5] Fang Wei, Yu Zhaofei, Chen Yanqi, et al. Incorporating learnable membrane time constant to enhance learning of spiking neural networks[C]//Proc of the 21st IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2021: 2661−2671

[6] Guo Yufei, Chen Yuanpei, Liu Xiaode, et al. Ternary spike: Learning ternary spikes for spiking neural networks[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024, 12244−12252

[7] Luo Xinhao, Yao Man, Chou Yuhong, et al. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection[C]//Proc of the 18th European Conf on Computer Vision (ECCV). Berlin: Springer, 2024: 253−272

[8] Hu Yifan, Deng Lei, Wu Yujie, et al. Advancing spiking neural networks toward deep residual learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(2): 2353−2367

[9] Su Qiaoyi, Chou Yuhong, Hu Yifan, et al. Deep directly-trained spiking neural networks for object detection[C]//Proc of the 22nd IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2023: 6555−6565

[10] Yao Man, Zhao Guangshe, Zhang Hengyu, et al. Attention spiking neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9393−9410

[11] Yang Shu, Yu Chengting, Liu Lei, et al. Efficient ANN-guided distillation: Aligning rate-based features of spiking neural networks through hybrid block-wise replacement[C]//Proc of the 42nd

IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2025: 10025−10035

[12] Yu Chengting, Zhao Xiaochen, Liu Lei, et al. Efficient logit-based knowledge distillation of deep spiking neural networks for full-range timestep deployment[C/OL]//Proc of the 42nd Int Conf on Machine Learning (ICML). 2025: 393−941. [2026-01-21]. https://openreview.net/forum?id=ZvkyeUrpsA

[13] Cao Yongqiang, Chen Yang, Khosla D. Spiking deep convolutional neural networks for energy-efficient object recognition[J]. International Journal of Computer Vision, 2015, 113(1): 54−66

[14] Hunsberger E, Eliasmith C. Spiking deep networks with LIF neurons[J]. arXiv preprint, arXiv: 1510.08829, 2015

[15] Hunsberger E, Eliasmith C. Training spiking deep networks for neuromorphic hardware[J]. arXiv preprint, arXiv: 1611.05141, 2016

[16] Diehl P U, Neil D, Binas J, et al. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing[C]//Proc of the 2015 Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2015: 1−8

[17] Zhang Anguo, Shi Jieming, Wu Junyi, et al. Low latency and sparse computing spiking neural networks with self-driven adaptive threshold plasticity[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(12): 17177−17188

[18] Bu Tong, Ding Jianhao, Yu Zhaofei, et al. Optimized potential initialization for low-latency spiking neural networks[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022, 11−20

[19] Yao Xingting, Li Fanrong, Mo Zitao, et al. GLIF: A unified gated leaky integrate-and-fire neuron for spiking neural networks[C]//Proc of the 36th Int Conf on Neural Information Processing Systems (NeurIPS). San Diego, CA: Neural Information Processing Systems Foundation, Inc, 2022: 32160−32171

[20] Guo Yufei, Liu Xiaode, Chen Yuanpei, et al. Spiking Transformer: introducing accurate addition-only spiking self-attention for transformer[C]//Proc of the 42nd IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2025: 24398−24408

[21] Guo Yufei, Zhang Liwen, Chen Yuanpei, et al. Real spike: Learning real-valued spikes for spiking neural networks[C]//Proc of the 17th European Conf on Computer Vision (ECCV). Berlin: Springer, 2022: 52−68

[22] Feng Shuo, Lu Dongdong, Yin Fei, et al. Instruction set design for brain-inspired intelligent processors oriented to distributed computing[J]. Journal of Computer Research and Development, 2025, 62(1): 1−14 (in Chinese)
(冯烁, 路冬冬, 尹飞, 等. 面向分布式计算的类脑智能处理器指令集设计[J]. 计算机研究与发展, 2025, 62(1): 1−14)

[23] Yang Zhijie, Wang Lei, Shi Wei, et al. Asynchronous network-on-chip architecture for brain-inspired processors[J]. Journal of Computer Research and Development, 2023, 60(1): 17−29 (in Chinese)
(杨智杰, 王蕾, 石伟, 等. 类脑处理器异步片上网络架构[J]. 计算机研究与发展, 2023, 60(1): 17−29)

[24] Wang Zhichao, Chen Liang, Li Qianpeng, et al. A brain-inspired network-on-chip architecture with hybrid-mode routing[J]. Journal of Computer Research and Development, 2025, 62(11): 2663−2676 (in Chinese)
(王智超, 陈亮, 李千鹏, 等. 基于混合模式路由的脑启发片上网络架构[J]. 计算机研究与发展, 2025, 62(11): 2663−2676)

[25] Sengupta A, Ye Yuting, Wang R, et al. Going deeper in spiking neural networks: VGG and residual architectures[J]. Frontiers in Neuroscience, 2019, 13(1): 95−106

[26] Han Bing, Srinivasan G, Roy K. RMP-SNN: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 13558−13567

[27] Yoo D, Jeong D S. Cbp-Qsnn: Spiking neural networks quantized using constrained backpropagation[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2023, 13(4): 1137−1146

[28] Meng Qingyan, Xiao Mingqing, Yan Shen, et al. Training high-performance low-latency spiking neural networks by differentiation on spike representation[C]//Proc of the 39th IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2022: 12444−12453

[29] Zhou Zhaokun, Zhu Yuesheng, He Chao, et al. SpikFormer: When spiking neural network meets transformer[C/OL]//Proc of the 11th Int Conf on Learning Representations (ICLR). 2023[2026-01-21]. https://openreview.net/forum?id=frE4fUwz_h

[30] Yao Man, Hu Jiakui, Zhou Zhaokun, et al. Spike-driven Transformer[C]//Proc of the 37th Int Conf on Neural Information Processing Systems (NeurIPS). New Orleans, CA: Curran Associates, 2023: 64043−64058

[31] Fang Yi, Wang Zhaokun, Zhang Lei, et al. Spiking wavelet Transformer[C]//Proc of the 18th European Conf on Computer Vision (ECCV). Berlin: Springer, 2024: 19−37

[32] Wang Shu, Zhang Ming, Zhang Di, et al. Spiking vision Transformer with saccadic attention[C/OL]//Proc of the Int Conf on Learning Representations (ICLR). 2025[2026-01-21]. https://openreview.net/forum?id=qzZsz6MuEq

[33] Liang Yixing, Wei Wenjie, Belatreche A, et al. Towards accurate binary spiking neural networks: Learning with adaptive gradient modulation mechanism[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025, 39(2): 1402−1410

[34] Zhang Tianpeng, Yu Kexin, Zhong Xinghao, et al. STAA-SNN: Spatial-temporal attention aggregator for spiking neural networks[C]//Proc of the 42nd IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2025: 13959−13969

[35] Deng Shikuang, Li Yang, Zhang Shiyan, et al. Temporal efficient training of spiking neural network via gradient re-weighting [C/OL]//Proc of the Int Conf on Learning Representations (ICLR). 2022[2026-01-21]. https://openreview.net/forum?id=_XNtisL32jv

[36] Guo Yufei, Tong Xinyi, Chen Yuanpei, et al. RecDis-SNN: Rectifying membrane potential distribution for directly training spiking neural networks[C]//Proc of the 39th IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2022: 326−335

[37] Guo Yufei, Liu Xiaode, Chen Yuanpei, et al. RMP-loss: Regularizing

membrane potential distribution for spiking neural networks[C]//Proc of the 22nd IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2023: 17391−17401

[38] Guo Yufei, Zhang Yuhan, Chen Yuanpei, et al. Membrane potential batch normalization for spiking neural networks[C]//Proc of the 22nd IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2023: 19420−19430

[39] Yu Kexin, Zhang Tianpeng, Wang Hao, et al. FSTA-SNN: Frequency-based spatial-temporal attention module for spiking neural networks[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 22227−22235

[40] Guo Yufei, Chen Yuanpei, Zhang Liwen, et al. IM-loss: Information maximization loss for spiking neural networks[C]//Proc of the 36th Int Conf on Neural Information Processing Systems (NeurIPS). San Diego, CA: Neural Information Processing Systems Foundation, Inc, 2022: 156−166



Wang Yufei, born in 1995. PhD. His main research interests include machine vision and neuromorphic computing.

王宇飞, 1995 年生。博士。主要研究方向为机器视觉、类脑计算。



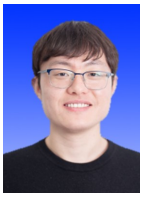
Zhao Jingyue, born in 1994. PhD, assistant professor. Member of CCF. Her main research interests include neuromorphic computing, brain-inspired robotics, and embodied intelligence.

赵静月, 1994 年生。博士，助理研究员。CCF 会员。主要研究方向为类脑计算、类脑机器人、具身智能。



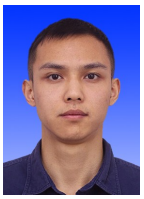
Zhang Ge, born in 1996. PhD, assistant professor. Her main research interests include hyperspectral image interpretation and multimodal image fusion.

张　歌, 1996 年生。博士，助理研究员。主要研究方向为高光谱图像解译、多模态图像融合。



Fan Bin, born in 1994. PhD, professor, PhD supervisor. His main research interests include machine vision, computational photography, and 3D reconstruction.

樊　斌, 1994 年生。博士，教授，博士生导师。主要研究方向为机器视觉、计算摄影、3 维重建。



Yang Zhijie, born in 1995. PhD. His main research interests include neuromorphic computing and computer architecture.

杨智杰, 1995 年生。博士。主要研究方向为类脑计算、计算机体系结构。



Wan Zhong, born in 1989. PhD, associate professor. Her main research interest includes neuromorphic computing.

万　众, 1989 年生。博士，副研究员。主要研究方向为类脑计算。



Dai Yuchao, born in 1982. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include machine vision and embodied intelligence.

戴玉超, 1982 年生。博士，教授，博士生导师。CCF 高级会员。主要研究方向为机器视觉、具身智能。



Wang Lei, born in 1977. PhD, professor, PhD supervisor. Distinguished member of CCF. Her main research interests include computer architecture and intelligent chips.

王　蕾, 1977 年生。博士，研究员，博士生导师。CCF 杰出会员。主要研究方向为计算机体系结构、智能芯片。