SpikePack: Enhanced Information Flow in Spiking Neural Networks with High Hardware Compatibility

Guobin Shen^{1,2}, Jindong Li^{1,3}, Tenglong Li^{1,3} Dongcheng Zhao¹, and Yi Zeng^{1,2,3,†}

¹ BrainCog Lab, CASIA, ² School of Future Technology, UCAS, ³ School of Artificial Intelligence, UCAS {shenguobin2021, lijindong2022, litenglong2023, zhaodongcheng2016, yi.zeng}@ia.ac.cn

Abstract

Spiking Neural Networks (SNNs) hold promise for energyefficient, biologically inspired computing. We identify substantial information loss during spike transmission, linked to temporal dependencies in traditional Leaky Integrateand-Fire (LIF) neurons—a key factor potentially limiting SNN performance. Existing SNN architectures also underutilize modern GPUs, constrained by single-bit spike storage and isolated weight-spike operations that restrict computational efficiency. We introduce SpikePack, a neuron model designed to reduce transmission loss while preserving essential features like membrane potential reset and leaky integration. SpikePack achieves constant O(1) time and space complexity, enabling efficient parallel processing on GPUs and also supporting serial inference on existing SNN hardware accelerators. Compatible with standard Artificial Neural Network (ANN) architectures, SpikePack facilitates near-lossless ANN-to-SNN conversion across various networks. Experimental results on tasks such as image classification, detection, and segmentation show SpikePack achieves significant gains in accuracy and efficiency for both directly trained and converted SNNs over state-of-theart models. Tests on FPGA-based platforms further confirm cross-platform flexibility, delivering high performance and enhanced sparsity. By enhancing information flow and rethinking SNN-ANN integration, SpikePack advances efficient SNN deployment across diverse hardware platforms.

1. Introduction

Spiking Neural Networks (SNNs) [32] have emerged as a promising paradigm for energy-efficient and biologically inspired computing [54]. By emulating the discrete spike-based communication of biological neurons, SNNs offer potential advantages in terms of low-power consumption and event-driven processing, which are particularly appealing for deployment on neuromorphic hardware [22–24, 34, 36].

Despite these advantages, SNNs still face significant

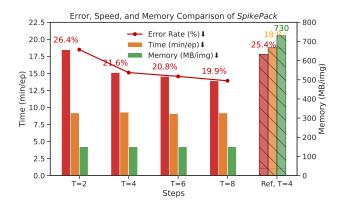


Figure 1. Performance of *SpikePack* on Spikeformer-8-512 across different time steps (T), showing error rate (%), time (ep/min), and memory (MB/img). Reference model uses LIF neurons (Ref, T=4).

challenges that impede their widespread application in complex tasks such as image classification [28, 38, 46, 55, 58], object detection [26, 30, 51], and natural language processing [39, 49]. Notably, their performance often lags behind that of Artificial Neural Networks (ANNs). One key reason we have identified is the substantial information loss that occurs during spike transmission, particularly associated with the temporal dependencies inherent in traditional neuron models like the Leaky Integrate-and-Fire (LIF) neuron [3]. This information degradation can limit the network's ability to capture and transmit critical features, thereby hindering overall performance.

Moreover, existing SNN architectures have not fully exploited the capabilities of modern General-Purpose Graphics Processing Units (GPGPUs). The reliance on single-bit spike representations [40] and isolated weight-spike operations [30, 41] leads to inefficient utilization of parallelism, resulting in low training efficiency and slow inference speeds. This inefficiency not only hampers the practicality of SNNs but also complicates their deployment across diverse hardware platforms [2, 20].

Although there has been considerable research into neuromorphic hardware [37] and the development of various

SNN accelerators [9, 23], SNNs have not become mainstream. This is partly due to their subpar performance compared to ANNs and their incompatibility with modern ANN architectures [8, 58]. This incompatibility arises not only from the reliance on discrete spikes but also from the temporal dependencies in spiking neurons, which require SNN-specific network designs and training methods. Consequently, adapting ANN models and techniques to SNNs requires complex modifications, limiting SNNs' ability to fully leverage advancements in ANN architectures and optimization.

To address these challenges, we propose SpikePack, a novel neuron model designed to reduce information loss during the transition from pre-synaptic to post-synaptic spikes, minimizing degradation associated with temporal dependencies in traditional neuron models. SpikePack enhances information flow within SNNs while achieving $\mathcal{O}(1)$ time and space complexity with respect to time steps, enabling efficient time-parallel training and inference on GPUs, as shown in Figure 1. SpikePack also preserves essential biological characteristics, such as membrane potential reset and leaky integration, ensuring biological plausibility.

In addition, *SpikePack* is compatible with modern ANN architectures, allowing for near-lossless ANN-to-SNN conversion and preserving the inherent sparsity of spike-based computations. This compatibility enables the integration of advanced ANN models within the SNN framework, improving performance across a variety of tasks.

Our contributions can be summarized as follows:

- We introduce *SpikePack*, a neuron model that minimizes information loss from pre-synaptic to post-synaptic spikes. *SpikePack* achieves a balance between computational efficiency and biological fidelity in SNNs.
- By achieving O(1) time and space complexity, SpikePack preserves essential neural dynamics, enabling efficient, biologically relevant behavior. As shown in Figure 2, SpikePack supports direct training, eliminating the need for complex temporal unfolding and enabling more efficient gradient-based optimization. Its compatibility with modern ANN architectures further supports near-lossless ANN-to-SNN conversion while maintaining the inherent sparsity of spike-based computations.
- Extensive experiments on image classification, object detection, and segmentation tasks showcase significant improvements over state-of-the-art methods. Additional testing on SNN hardware accelerators further validates the generality and efficiency of our approach.

By addressing the fundamental issues of information loss and hardware inefficiency, *SpikePack* represents a significant advancement in the practical deployment of SNNs. It not only enhances the computational capabilities of SNNs but also ensures that these improvements are accessible across various hardware platforms. This work not only brings us closer to realizing the full potential of neuromorphic computing in real-world scenarios but also offers new insights into bridging SNNs and ANNs.

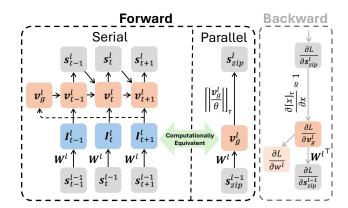


Figure 2. Forward and backward computation in SpikePack, showing both serial and parallel modes. Parallel computation uses $s_{\rm zip}^l$ for efficient global potential (v_g^l) calculation.

2. Related Work

Our work builds on advancements in three areas of SNNs: information transmission efficiency, efficient training methods, and compatibility with ANNs.

Information Transmission Efficiency Traditional neuron models like the LIF neuron rely on binary spikes, which limit information capacity. To address this, ternary spike neurons [12] and burst-based LIFB neurons [40] expand spike representations, allowing richer information flow while retaining energy efficiency. Other approaches, such as integer-valued neurons [30] and rectified membrane potentials [11], aim to reduce quantization errors and mitigate degradation in deep SNNs. Although these methods improve information transmission, they often add model complexity. Our *SpikePack* offers a simpler approach by computing membrane potentials before spikes are generated, effectively enhancing information flow without increasing computational demands.

Efficient Training Methods Training efficiency is a core challenge in SNNs due to the non-differentiability of spike operations, and weight-spike computations, all of which limit hardware utilization. While Backpropagation Through Time (BPTT) [46] has enabled deep SNN training, it incurs high memory and time costs. Approaches like Online Training Through Time (OTTT) [47] and Spatial Learning Through Time (SLTT) [33] reduce memory overhead by prioritizing critical temporal interactions, while Temporal Reversible SNNs (T-RevSNN) [14] leverage reversible architectures to lower memory costs. However, these methods primarily optimize training efficiency without addressing the fundamental limitations of binary spikes and underuti-

lized hardware. In contrast, SpikePack inherently supports time-parallel processing, achieving $\mathcal{O}(1)$ complexity with respect to time steps, thus improving training efficiency and enabling high inference efficiency across diverse hardware.

Compatibility with ANN Architectures Leveraging ANN advancements within the SNN framework is challenging due to fundamental differences in activation dynamics. Traditional ANN-to-SNN conversion approaches rely on rate coding, often requiring many time steps and recalibration [26]. Recently, methods such as Spatio-Temporal Approximation [17] and Expectation Compensation [16] have enabled SNN adaptations of Transformer architectures by approximating non-linear interactions. While effective, these methods are often complex and architecturespecific. By contrast, *SpikePack* enables near-lossless conversion across various ANN architectures with minimal adjustments, facilitating direct integration with modern ANN models while preserving SNN sparsity and efficiency.

While previous works have addressed aspects of information flow, training efficiency, and ANN compatibility, they often require complex modifications. *SpikePack* provides a unified, streamlined solution that enhances information flow, supports efficient training, and enables seamless integration with ANN architectures, promoting scalable SNN deployment across diverse hardware.

3. Methodology

In this section, we analyze the limitations of LIF neurons, focusing on their information transmission inefficiencies and computational limitations on GPGPUs. These limitations stem from both information loss during spike transmission and inefficient hardware utilization. To address these challenges, we propose *SpikePack*, a novel neuron model designed to preserve critical information and support efficient parallel processing on modern hardware. We also provide a theoretical foundation for *SpikePack* through analysis of information transmission, followed by an explanation of its computational efficiency.

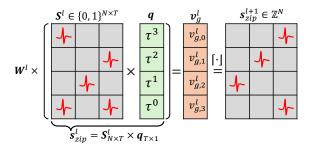


Figure 3. Spike sequence compression into integer s_{zip}^l for efficient computation of v_q^l without decompression.

3.1. Limitations of LIF Neurons

The LIF neuron model is widely used in SNNs due to its simplicity and biological inspiration. However, it suffers from two key limitations that hinder its effectiveness: low information retention during spike transmission and inefficient utilization of modern parallel processing hardware.

3.1.1. Low Information Capacity in Spike Transmission

LIF neurons generate output spikes based on the membrane potential at each discrete time step. However, since the membrane potential is determined by the combined effects of the historical input sequence and the decay factor, the model exhibits certain limitations in fully integrating input information. This means that each spiking decision is based on an incomplete view of the input sequence, causing substantial information loss.

For a given layer l in an SNN, let $\mathbf{S}^{l-1} \in \{0,1\}^{N \times T}$ denote the binary spike matrix from the previous layer, where N is the number of pre-synaptic neurons, and T is the number of time steps. The weight matrix of layer l is represented as $\mathbf{W}^l \in \mathbb{R}^{M \times N}$, where M is the number of neurons in the current layer. For our analysis, we focus on a single neuron in this layer, represented by the weight vector $\mathbf{w}^l = \mathbf{W}^l_{i,..}$ Here, we refer to the output spike sequence for this neuron as $\mathbf{s}^l \in \{0,1\}^{1 \times T}$.

The membrane potential v_t of a LIF neuron updates according to the rule given in Eq. (1):

$$v_{t}^{l} = \frac{1}{\tau} v_{t-1}^{l} + \mathbf{w}^{l} \cdot \mathbf{s}_{t}^{l-1} - \theta \cdot s_{t-1}^{l}, \tag{1}$$

where τ is the membrane time constant, \mathbf{s}_t^{l-1} represents input spikes at time t, and θ denotes the firing threshold. A spike is generated when v_t exceeds the threshold θ , after which the membrane potential is reset to facilitate subsequent spiking dynamics as shown in Eq. (2):

$$s_t^l = \begin{cases} 1, & \text{if } v_t > \theta, \\ 0, & \text{otherwise.} \end{cases}$$
 (2)

Since v_t^l is recursively dependent on v_{t-1}^l , each spiking decision is based on partially integrated information from the input sequence. This recursive structure imposes inherent limitations on the mutual information that can be preserved between the input and output, leading to significant information loss during spike transmission. As a result, the ability of the model to fully encode and utilize temporal dependencies in the input signal is constrained.

3.1.2. Inefficient Hardware Utilization

Another significant limitation of LIF neurons lies in their inefficient utilization of modern parallel processing hardware. The recursive nature of the membrane potential update, as defined in Eq. (1), inherently restricts parallelism, as each

time step must be processed sequentially, resulting in a time complexity of $\mathcal{O}(T)$, where T is the number of time steps. Furthermore, the need to store individual spikes for each time step occupies separate integer or floating-point units in memory, leading to inefficient memory utilization and increased space complexity of $\mathcal{O}(T)$. These inefficiencies are especially problematic for hardware architectures optimized for parallel computation.

In addition, the LIF model requires repeated spike-weight operations at every time step. As each input at each time step consists of a single spike value, the effective computation is limited to spike-weight multiplications, thereby underutilizing the General Matrix Multiplication (GeMM) capabilities of modern hardware. During training, the reliance on surrogate gradient approximations [46] to handle the non-differentiable spiking functions further exacerbates computational inefficiency, increasing both the computational overhead and overall complexity.

3.2. SpikePack

To address these challenges, we introduce SpikePack, a neuron model designed to preserve information capacity across the input sequence while supporting efficient parallel computation. SpikePack aggregates information across the entire input sequence into a global membrane potential, denoted as v_g^l for layer l. This global aggregation enhances information flow and maximizes mutual information between inputs and outputs.

The global membrane potential v_q^l is computed as:

$$v_q^l = \mathbf{w}^l \mathbf{S}^{l-1} \mathbf{q},\tag{3}$$

where $\mathbf{q} = [\tau^{T-1}, \tau^{T-2}, \dots, \tau^0]^{\top}$ applies the influence of the membrane time constant τ across time steps. This formulation enables v_g^l to integrate information from the entire input sequence \mathbf{S}^{l-1} , resulting in improved information flow from pre-synaptic to post-synaptic neurons.

After aggregating information into v_g^l , SpikePack generates the output spike sequence \mathbf{S}^l through a decoding process. The membrane potential v_t^l is updated as in Eq. (4):

$$v_t^l = v_{t-1}^l - \theta_t \cdot s_{t-1}^l, \tag{4}$$

where $\theta_t = \frac{\theta}{\tau^{t-T}}$ represents a dynamic threshold that adapts over time. The initial membrane potential is set to $v_0^l = v_g^l$. The spike generation condition is defined as in Eq. (5):

$$s_t^l = \begin{cases} 1, & \text{if } v_t^l > \theta_t, \\ 0, & \text{otherwise.} \end{cases}$$
 (5)

3.2.1. Improved Information Capacity in SpikePack

The initial global membrane potential v_g^l aggregates information across the entire input sequence, thereby enhancing the mutual information between the input \mathbf{S}^{l-1} and the

output spike sequence \mathbf{s}^l . Assuming binary spikes with independent Bernoulli distributions and Gaussian weights, v_g^l approximates a Gaussian distribution with variance as shown in Eq. (6):

$$\sigma_{v_g^l}^2 = \sigma^2 N p (1 - p) \left(\sum_{t=1}^T q_t\right)^2,$$
 (6)

where p represents the probability of an input spike, and $q_t = \tau^{t-1}$.

The mutual information between S^{l-1} and s^l for SpikePack can thus be approximated as in Eq. (7):

$$I_{\rm SP}^l = \frac{1}{2} \log_2 \left(\frac{12\sigma_{v_g^l}^2}{\theta^2} \right).$$
 (7)

This result, derived theoretically and validated empirically, demonstrates that SpikePack achieves greater information retention across a range of configurations for N and T, affirming its superior transmission capacity.

For a more detailed theoretical derivation of the mutual information $I(\mathbf{S}_{\text{in}}, \mathbf{s}_{\text{out}})$ between pre-synaptic and post-synaptic spikes in both the LIF and SpikePack models, please refer to Appendix A. Empirical simulations, as illustrated in Figure 4, confirm that SpikePack consistently achieves higher mutual information across various values of N and T, further substantiating its enhanced information transmission capability.

3.2.2. Parallel Computation and Hardware Utilization

The SpikePack model leverages the compressed input structure v_g^l , making its operations hardware-friendly and well-suited for parallel computation. As shown in Figure 3, since

Empirical Mutual Information SpikePack LIF Neuron

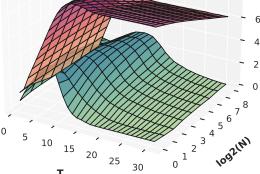


Figure 4. Empirical comparison of mutual information between SpikePack and LIF neurons over varying T (time steps) and N (pre-synaptic neurons). SpikePack demonstrates higher information retention across configurations.

all spike information is aggregated into v_g^l , both time and space complexity are reduced from $\mathcal{O}(T)$ to $\mathcal{O}(1)$, enabling efficient use of parallel processing on modern GPUs.

The packed representation of input spikes enables highly efficient matrix multiplication operations, thereby optimizing computational performance. The global membrane potential v_q^l is computed as in Eq. (8):

$$v_q^l = \mathbf{w}^l \mathbf{s}_{\text{zip}}^{l-1},\tag{8}$$

where $\mathbf{s}_{\mathrm{zip}}^{l-1} = \mathbf{S}^{l-1}\mathbf{q}$. This formulation compresses the presynaptic spike matrix into a bitwise representation, facilitating optimized GeMM operations and enhancing computational efficiency.

Utilizing this compressed representation, the total output spike count over T time steps can be derived directly from v_a^l without iterative updates, as in Eq. (9):

$$s_{\rm zip}^l = \left\lceil \frac{v_g^l}{\theta} \right\rceil_{\tau}.$$
 (9)

In this representation, each bit in s_{zip}^l indicates whether an output spike s_t^l is generated at time step t, effectively compressing the spike sequence. This compact form aligns with the serial spike generation in Eqs. (4) and (5), preserving spiking behavior without per-step computations and significantly reducing computational and memory demands.

By treating the membrane time constant τ as a form of quantization, SpikePack adapts to various hardware configurations, balancing precision and efficiency. For instance, $\tau=2$ results in uniform quantization, while $\tau\neq 2$ enables non-uniform quantization, allowing the model to adjust its computational footprint based on hardware constraints.

3.2.3. Efficient Gradient Propagation

As shown in Figure 2, SpikePack enhances computational efficiency by simplifying gradient propagation, eliminating the need for BPTT. Instead, gradients are computed directly with respect to the compressed input structure \mathbf{s}_{zip}^{l-1} , which encapsulates the entire sequence of input spikes in a single compressed representation. This method significantly reduces memory consumption and computational complexity during training by removing the requirement to unroll across time s

The gradient of the loss \mathcal{L} with respect to $\mathbf{s}_{\mathrm{zip}}^{l-1}$ is computed as shown in Eq. (10):

$$\frac{\partial \mathcal{L}}{\partial \mathbf{s}_{zip}^{l-1}} = \frac{\partial \mathcal{L}}{\partial \mathbf{s}_{zip}^{l}} \frac{\partial \mathbf{s}_{zip}^{l}}{\partial v_{q}^{l}} \frac{\partial v_{g}^{l}}{\partial \mathbf{s}_{zip}^{l-1}} \approx \frac{\partial \mathcal{L}}{\partial \mathbf{s}_{zip}^{l}} \frac{\mathbf{w}^{l}}{\theta}, \quad (10)$$

where we approximate $\frac{\partial \lceil x \rfloor_{\tau}}{\partial x} \approx 1$, effectively bypassing the non-differentiable quantization step. This direct gradient path allows for efficient, memory-saving training

aligned with SpikePack's compressed and parallel computation model.

By leveraging this streamlined gradient computation, *SpikePack* enables gradient propagation without costly temporal dependencies, as required in traditional SNN models. This design is inherently compatible with hardware architectures that support vectorized and parallel computation, such as SIMD instructions and systolic arrays.

4. Experiment

We evaluate *SpikePack* on tasks including image classification, object detection, and semantic segmentation, comparing its performance with state-of-the-art SNN models, neuron designs, and ANN-to-SNN conversion methods. We also conduct ablation studies to assess the impact of key parameters, demonstrating the versatility and efficiency of *SpikePack* across various datasets.

4.1. Experimental Setup

We conduct experiments on both static and neuromorphic datasets to thoroughly assess our model. For image classification, we use ImageNet dataset [4]; for object detection, the COCO 2017 dataset [27]; and for semantic segmentation, the ADE20K dataset [57]. To evaluate event-based performance, we use neuromorphic datasets including CIFAR10-DVS [21], DVS-Gesture [1], and N-Caltech101 [35]. Experiments are implemented in PyTorch and run on NVIDIA A100 GPUs, with a default membrane time constant of $\tau=2$. Additional experimental details are provided in Appendix B.

4.2. Image and Neuromorphic Data Classification

We evaluate *SpikePack* on the ImageNet dataset and compare its performance with other SNN models and neuron designs.

Comparison with Other Neuron Models We benchmark *SpikePack* against several neuron models, including LIF, LIFB [40], PSN [10], DSGM [42], and GLIF [52]. For fair evaluation, experiments are conducted using the SEW-ResNet [8] and Spikeformer [59] architectures. Model scales and time steps are adjusted to achieve comparable performance metrics across setups.

Figure 5 illustrates the efficiency of SpikePack, showcasing its balance between accuracy and computational cost, quantified by Synaptic Operations (SOPs). SOPs measure the overall spiking activity and computational workload, defined as the product of the firing rate, operations per time step, and the number of time steps T.

Compared to other neuron models, *SpikePack* consistently achieves higher accuracy at similar or lower computational costs. For instance, within the Spikeformer architecture, *SpikePack* matches the accuracy of competing neuron

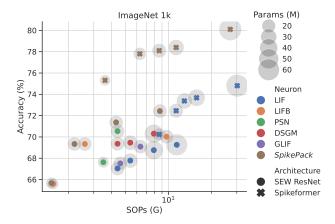


Figure 5. Comparison between *SpikePack* and other neuron models at different time steps on ImageNet 1k. Our method achieves higher accuracy with lower computational cost.

Table 1. Comparison of ANN-to-SNN conversion methods on ImageNet 1k. We report the top-1 accuracy (%) at different time steps (T).

Model	Method	ANN	T=32	T=64	T=128	T=256
	RMP [13]	-	-	-	-	55.65
DagNat 24	Opt. [5] Calib. [25]	70.64	33.01	59.52	67.54	70.06
Resnet-34	Calib. [25]	70.95	64.54	71.12	73.45	74.61
	SNM [44]	75.66	55.28	62.72	65.53	69.31
ViT-B/32	STA [17]	73.30	78.72	82.33	82.56	82.79
			T=5	T=6	T=8	T=12
ResNet-34	G . 1 . D 1	80.72	59.32	74.68	77.81	77.92
ViT-B/32	<i>зріке</i> гаск	77.92	57.23	79.26	80.68	80.72
ViT-L/14	Ours	88.27	85.59	88.00	88.22	88.27

models while requiring only 1/10 of the SOPs. Furthermore, at equivalent SOP levels, SpikePack delivers nearly a 5% improvement in accuracy, highlighting its superior information transmission and computational efficiency. This advantage persists across varying model sizes and architectures, demonstrating the robustness and scalability of SpikePack.

Comparison with Efficient SNN Training Methods We evaluate *SpikePack* against several efficient SNN training methods, including STBP-tdBN [56], SEW ResNet [8], MS ResNet [15], TEBN [7], TET [6], OTTT [47], SLTT [33], Parallel SNN [10], and T-RevSNN [14].

As shown in Table 2, *SpikePack* outperforms these methods in accuracy while requiring less training time and memory. For instance, using ResNet-34 with 4 time steps, *SpikePack* achieves 73.4% accuracy with only 8.1 minutes per epoch and 24.3 MB memory per image, demonstrating reduced training overhead and superior performance. Moreover, experiments with Transformer-based architectures, such as Spikeformer, confirm that *SpikePack* main-

tains $\mathcal{O}(1)$ time and space complexity relative to the number of time steps T, achieving up to 80.1% accuracy with 8 time steps without any increase in memory or computational load as T grows.

Comparison with ANN-to-SNN Conversion Methods We evaluate *SpikePack* for near-lossless ANN-to-SNN conversion, leveraging its compatibility with various ANN architectures as discussed in Section 3.2.2. Unlike other methods, *SpikePack* enables efficient conversion without requiring post-conversion training or calibration.

As shown in Table 1, we compare *SpikePack* with several conversion methods, including RMP [13], Optimal (Opt.) [5], Spike Calibration (Cailb.) [25], SNM [44], and Spatio-Temporal Approximation (STA) [17]. *SpikePack* achieves high accuracy with as few as 6 time steps and nearlossless performance at 8 time steps—less than 1/10 of the steps required by other methods. For example, *SpikePack* achieves 77.92% accuracy with ResNet-34 and 88.27% with ViT-L/14, closely matching ANN performance. Figure 6 further highlights *SpikePack*'s ability to maintain high accuracy across different architectures with minimal time steps compared to competing methods.

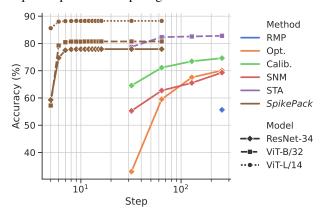


Figure 6. Comparison of ANN-to-SNN conversion methods at different time steps. *SpikePack* achieves higher accuracy with fewer time steps.

Results on Neuromorphic Datasets We evaluate *SpikePack* on neuromorphic datasets including CIFAR10-DVS [21], DVS-Gesture [1], and N-Caltech101 [35]. Table 3 reports the accuracy at different time steps.

SpikePack achieves competitive accuracy on neuromorphic datasets, demonstrating adaptability to event-based data. Unlike static datasets, neuromorphic datasets generally require a higher number of time steps T for optimal performance—a trend consistent with prior studies [41]. These results confirm SpikePack's versatility in effectively handling both static and event-driven data.

4.3. Object Detection

We evaluate *SpikePack* on the COCO 2017 validation dataset [27], leveraging RTMDet [31] and DINO [53] as

Table 2. Comparison of SpikePack with state-of-the-art SNN training methods on ImageNet 1k. We report the number of parameters, time
steps, training time per epoch, memory usage per image, synaptic operations (SOPs), and top-1 accuracy.

Maderal	A 1. 24 4	Param	Time	Training Time	Memory	SOP	A (01)
Methods	Architecture	(M)	steps	(min/ep)	(MB/img)	(G)	Acc (%)
STBP-tdBN [56]	ResNet-34	21.8	6	29.6	186.1	7.1	63.7
SEW ResNet [8]	SEW-ResNet-34	21.8	4	5.0	224.5	4.4	67.0
SEW RESNET [6]	SEW-ResNet-50	25.6	4	10.0	596.9	5.4	67.8
MS ResNet [15]	MS-ResNet-34	21.8	6	11.2	267.1	5.7	69.4
TEBN [7]	ResNet-34	21.8	4	16.3	260.1	7.1	64.3
TET [6]	SEW-ResNet-34	21.8	4	12.5	221.0	4.4	68.0
Spikformer [59]	Spikeformer-8-384	16.8	4	14.2	580.8	8.6	70.2
Spikionnei [39]	Spikeformer-8-512	29.7	4	16.7	767.8	12.9	73.4
Spike-driven	Spikeformer-8-384	16.8	4	15.4	548.9	4.3	72.3
Transformer [50]	Spikeformer-8-512	29.7	4	18.8	730.0	5.1	74.6
OTTT [47]	ResNet-34	21.8	6	24.2	84.1	6.7	64.2
SLTT [33]	ResNet-34	21.8	6	18.1	71.7	6.7	66.2
SLI1 [33]	ResNet-50	25.6	6	23.4	117.3	8.0	67.0
Parallel SNN [10]	SEW-ResNet-18	11.7	4	5.8	138.7	-	67.6
raraner Siviv [10]	SEW-ResNet-34	21.8	4	8.3	179.7	4.4	70.5
T-RevSNN [14]	ResNet-18	15.2	4	6.1	57.5	1.9	69.8
1-Kev5MN [14]	ResNet-18	29.8	4	9.1	85.7	3.1	73.2
	ResNet-18	11.1	4	5.8	20.1	1.8	70.6
	ResNet-34	21.8	4	8.1	24.3	3.7	73.4
	ResNet-50	25.6	4	13.5	53.4	4.1	78.7
SpikePack (Ours)	Spikeformer-8-512	29.7	2	9.2	150.5	3.9	73.6
	Spikeformer-8-512	29.7	4	9.2	150.5	7.7	78.4
	Spikeformer-8-512	29.7	6	9.2	150.5	11.2	79.2
	Spikeformer-8-512	29.7	8	9.2	150.5	15.1	80.1

Table 3. Classification accuracy on neuromorphic datasets at different time steps ${\cal T}.$

Dataset	T=4	T=6	T=8	T=10	T=12	T=16
CIFAR10-DVS [21] DVS-Gesture [1] N-Caltech101 [35]	68.1	76.4	80.7	82.4	83.7	84.6
DVS-Gesture [1]	94.6	95.5	96.7	97.4	96.7	97.4
N-Caltech101 [35]	76.3	80.0	81.7	82.2	82.5	82.6

our base architectures. To highlight the effectiveness of our approach, we compare it against other SNN-based object detection models, including Spiking-YOLO [19], Bayesian Optimization [18], Spike Calibration [25], EMS-YOLO [43], Meta-SpikeFormer [51], and SpikeYOLO [30].

Table 4 summarizes the evaluation results. Our approach demonstrates superior mean Average Precision (mAP) while requiring fewer time steps and reducing computational cost. For instance, with DINO-r50 and just 6 time steps, *SpikePack* achieves an impressive 48.5% mAP@50:95, surpassing previous methods with a substantial reduction in computational overhead.

Our method demonstrates competitive performance with significantly fewer parameters and computational cost. For instance, using RTMDet-m with only 67.2G SOPs, we achieve 49.1% mAP@50:95 at 8 time steps.

4.4. Semantic Segmentation

Table 5 presents the segmentation results on ADE20K, showcasing *SpikePack*'s strong performance in dense prediction tasks like semantic segmentation. With Segformerb2 and 10 time steps, *SpikePack* achieves 45.6% mIoU, outperforming prior methods while significantly reducing computational cost. These results emphasize the method's efficiency and scalability for challenging benchmarks.

4.5. Ablation Studies

We conduct ablation studies to analyze the impact of the membrane time constant τ on SpikePack's performance, evaluated on ImageNet with ResNet-34 (Figure 7). While $\tau=4.0$ achieves the best performance by aligning with activation distributions, $\tau=2.0$ is more efficient on GPGPUs, avoiding exponentiation. As a result, $\tau=2.0$ has been the preferred choice in prior implementations.

4.6. Hardware Compatibility

Compatibility with Neuromorphic Processors As a spiking neuron, the *SpikePack* neuron adheres to the binary nature of spiking neurons, generating discrete spike sequences that are fully compatible with existing neuromorphic processors. With higher sparsity, *SpikePack* achieves superior

Table 4. Performance of object detection on COCO 2017 validation set [27]. We report the number of parameters, computational cost (SOPs), time steps, and mean Average Precision (mAP).

Model	Param (M)	SOPs (G)	Step	mAP@ 50(%)	mAP@ 50:95(%)
Spiking-YOLO [19]	10.2	-	3500	-	25.7
Bayesian Optim [18]	10.2	-	5000	-	25.9
Spike Calib [25]	17.1	-	512	45.4	-
EMS-YOLO[43]	26.9	32.2	4	50.1	30.1
Meta-SpikeFormer	34.9	55.0	1	44.0	-
(MaskRCNN) [51]	75.0	156.4	1	51.2	-
Meta-SpikeFormer	16.8	38.7	1	45.0	-
(YOLO) [51]	16.8	78.6	4	50.3	-
	23.1	38.6	4	62.3	45.5
SpikeYOLO [30]	48.1	76.1	4	64.6	47.4
	68.8	93.6	4	66.2	48.9
Sniko Dack (Ours)	4.8	8.53	6	55.7	39.0
SpikePack (Ours) w/ RTMDet-tiny [31]	4.8	11.3	8	57.8	40.9
w/ KTMDet-unly [31]	4.8	14.1	10	57.9	41.1
SpikePack (Ours)	24.7	51.8	6	50.1	48.5
w/ RTMDet-m [31]	24.7	67.2	8	61.7	49.1
w/ KTMDet-III [31]	24.7	86.3	10	61.9	49.4
Snika Daak (Oura)	47.7	276	6	66.0	48.5
SpikePack (Ours)	47.7	359	8	66.7	50.0
w/ DINO-r50 [53]	47.7	447	10	67.9	50.1

Table 5. Performance of semantic segmentation on ADE20K [57]. We report the number of parameters, computational cost (SOPs), simulation time steps, and mIoU (%).

Model	Param (M)	SOPs (G)	Step	MIoU(%) 50(%)
	16.5	24.6	1	32.3
Mata Spiles Former [51]	16.5	98.2	4	33.6
Meta-SpikeFormer [51]	58.9	51.7	1	34.8
	58.9	204.1	4	35.3
SnikoPack (Ours)	47.2	256.8	6	34.1
SpikePack (Ours) w/ FCN-r50	47.2	384.4	8	35.3
W/ FCN-130	47.2	476.7	10	35.9
SpikePack (Ours)	3.75	39.3	6	35.3
• '	3.75	51.7	8	36.9
w/ Segformer-b0 [48]	3.75	63.9	10	37.4
SpikePack (Ours)	24.8	83.6	6	42.8
• '	24.8	111.5	8	44.1
w/ Segformer-b2 [48]	24.8	138.7	10	45.6

Table 6. Latency and Energy Comparison between SpikePack and LIF neuron

		Pack	LIF		
	ResNet-34	ResNet-50	ResNet-34	ResNet-50	
Latency	23ms	24.1 ms	29.1 ms	34.7 ms	
Energy	18.6mJ	19.4 mJ	23.8 mJ	28.4 mJ	

speedup and reduced energy consumption compared to tra-

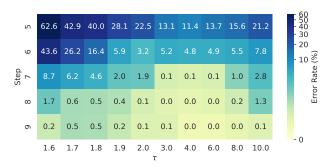


Figure 7. Ablation study on the effect of membrane time constant τ on ImageNet classification error rate (%) using ResNet-34 with *SpikePack*.

ditional LIF neurons. To validate this, we conduct hardware experiments comparing SpikePack and LIF neurons under identical conditions. We design a neuromorphic processor that processes binary spike inputs and synaptic weights, filtering zero elements in the spike tensor to ensure only active spikes contribute to computation. The processor includes 64 processing elements for synaptic current accumulation, 16 neuron dynamic units, and a 16-input spike detector that converts spike sequences into addresses for synaptic weight fetching. We implement the processor on xczu3eg FPGA chip running at 300MHz. The detailed implementation is shown in Appendix C. Using cycle-accurate simulation, we evaluate inference latency for ResNet-34 and ResNet-50 models. The results demonstrate that SpikePack achieves significantly lower latency and power consumption than traditional LIF neurons, highlighting its efficiency on neuromorphic hardware.

Compatibility with Parallel Computing Processors Modern parallel computing processors, such as GPGPUs, NPUs, and SIMD-enabled CPUs, excel at matrix multiplication—a critical operation in neural network acceleration. The SpikePack neuron efficiently utilizes these architectures by maintaining $\mathcal{O}(1)$ computation and memory usage across time steps. In contrast, traditional LIF neurons require $\mathcal{O}(T)$ duplication of computational workload and additional storage for membrane potentials, resulting in increased computational overhead.

4.7. Discussion

Our experimental results demonstrate that *SpikePack* significantly improves information flow and computational efficiency in SNNs. By minimizing information loss during spike transmission and supporting efficient parallel computation, *SpikePack* delivers superior performance across diverse tasks and datasets. Additionally, its seamless compatibility with standard ANN architectures enables nearlossless ANN-to-SNN conversion, allowing SNNs to benefit from the latest advancements in ANN models and training techniques.

References

- [1] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7243–7252, 2017. 5, 6, 7, 3
- [2] A. Carpegna, A. Savino, and S. D. Carlo. Spiker+: A framework for the generation of efficient spiking neural networks fpga accelerators for inference at the edge. *IEEE Transactions on Emerging Topics in Computing*, 01(01):1–15, 2024.
- [3] Peter Dayan and Laurence F Abbott. Theoretical neuroscience: computational and mathematical modeling of neural systems. MIT press, 2005. 1
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 5, 3
- [5] Shikuang Deng and Shi Gu. Optimal conversion of conventional artificial neural networks to spiking neural networks. In *International Conference on Learning Representations*, 2021. 6
- [6] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. arXiv preprint arXiv:2202.11946, 2022.
- [7] Chaoteng Duan, Jianhao Ding, Shiyan Chen, Zhaofei Yu, and Tiejun Huang. Temporal effective batch normalization in spiking neural networks. *Advances in Neural Information Processing Systems*, 35:34377–34390, 2022. 6, 7
- [8] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. Advances in Neural Information Processing Systems, 34:21056–21069, 2021. 2, 5, 6, 7
- [9] Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothée Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40):eadi1480, 2023. 2
- [10] Wei Fang, Zhaofei Yu, Zhaokun Zhou, Ding Chen, Yanqi Chen, Zhengyu Ma, Timothée Masquelier, and Yonghong Tian. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. Advances in Neural Information Processing Systems, 36, 2024. 5, 6, 7
- [11] Yufei Guo, Xinyi Tong, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Zhe Ma, and Xuhui Huang. Recdis-snn: Rectifying membrane potential distribution for directly training spiking neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 326–335, 2022. 2
- [12] Yufei Guo, Yuanpei Chen, Xiaode Liu, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. Ternary spike: Learning ternary spikes for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12244–12252, 2024. 2

- [13] Bing Han and Kaushik Roy. Deep spiking neural network: Energy efficiency through time based coding. In *European conference on computer vision*, pages 388–404. Springer, 2020. 6
- [14] JiaKui Hu, Man Yao, Xuerui Qiu, Yuhong Chou, Yuxuan Cai, Ning Qiao, Yonghong Tian, XU Bo, and Guoqi Li. High-performance temporal reversible spiking neural networks with o(l) training memory and o(1) inference cost. In Forty-first International Conference on Machine Learning, 2024. 2, 6, 7
- [15] Yifan Hu, Lei Deng, Yujie Wu, Man Yao, and Guoqi Li. Advancing spiking neural networks toward deep residual learning. IEEE Transactions on Neural Networks and Learning Systems, 2024. 6, 7
- [16] Zihan Huang, Xinyu Shi, Zecheng Hao, Tong Bu, Jian-hao Ding, Zhaofei Yu, and Tiejun Huang. Towards high-performance spiking transformers from ann to snn conversion. In ACM Multimedia 2024, 2024. 3
- [17] Yizhou Jiang, Kunlin Hu, Tianren Zhang, Haichuan Gao, Yuqian Liu, Ying Fang, and Feng Chen. Spatio-temporal approximation: A training-free snn conversion for transformers. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 6
- [18] Seijoon Kim, Seongsik Park, Byunggook Na, Jongwan Kim, and Sungroh Yoon. Towards fast and accurate object detection in bio-inspired spiking neural networks through bayesian optimization. *IEEE Access*, 9:2633–2643, 2020. 7, 8
- [19] Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: spiking neural network for energyefficient object detection. In *Proceedings of the AAAI con*ference on artificial intelligence, pages 11270–11277, 2020. 7. 8
- [20] Takumi Kuwahara, Reon Oshio, Mutsumi Kimura, Renyuan Zhang, and Yasuhiko Nakashima. Fusion synapse by memristor and capacitor for spiking neuromorphic systems. *Neu*rocomputing, 593:127792, 2024. 1
- [21] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017. 5, 6, 7, 3
- [22] Jindong Li, Guobin Shen, Dongcheng Zhao, Qian Zhang, and Yi Zeng. Firefly: A high-throughput hardware accelerator for spiking neural networks with efficient dsp and memory optimization. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2023. 1
- [23] Jindong Li, Guobin Shen, Dongcheng Zhao, Qian Zhang, and Yi Zeng. Firefly v2: Advancing hardware support for high-performance spiking neural network with a spatiotemporal fpga accelerator. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024. 2
- [24] Tenglong Li, Jindong Li, Guobin Shen, Dongcheng Zhao, Qian Zhang, and Yi Zeng. Firefly-s: Exploiting dual-side sparsity for spiking neural networks acceleration with reconfigurable spatial architecture. *IEEE Transactions on Circuits* and Systems I: Regular Papers, 2024. 1, 4
- [25] Yang Li and Yi Zeng. Efficient and accurate conversion of

- spiking neural network with burst spikes. arXiv preprint arXiv:2204.13271, 2022. 6, 7, 8
- [26] Yang Li, Xiang He, Yiting Dong, Qingqun Kong, and Yi Zeng. Spike calibration: Fast and accurate conversion of spiking neural network for object detection and segmentation. *arXiv preprint arXiv:2207.02702*, 2022. 1, 3
- [27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 5, 6, 8, 3
- [28] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018. 1
- [29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Pro*ceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015. 8
- [30] Xinhao Luo, Man Yao, Yuhong Chou, Bo Xu, and Guoqi Li. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. *arXiv preprint arXiv:2407.20708*, 2024. 1, 2, 7, 8
- [31] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 6, 8
- [32] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10 (9):1659–1671, 1997. 1
- [33] Qingyan Meng, Mingqing Xiao, Shen Yan, Yisen Wang, Zhouchen Lin, and Zhi-Quan Luo. Towards memory-and time-efficient backpropagation for training spiking neural networks. In *Proceedings of the IEEE/CVF International* Conference on Computer Vision, pages 6166–6176, 2023. 2, 6.7
- [34] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673, 2014. 1
- [35] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 5, 6, 7, 3
- [36] Sandeep Pande, Fearghal Morgan, Seamus Cawley, Brian McGinley, Snaider Carrillo, Jim Harkin, and Liam McDaid. Embrace-sysc for analysis of noc-based spiking neural network architectures. In 2010 International Symposium on System on Chip, pages 139–145. IEEE, 2010. 1
- [37] Jing Pei, Lei Deng, Sen Song, Mingguo Zhao, Youhui Zhang, Shuang Wu, Guanrui Wang, Zhe Zou, Zhenzhi Wu, Wei He, et al. Towards artificial general intelligence with hybrid tianjic chip architecture. *Nature*, 572(7767):106–111, 2019.

- [38] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Backpropagation with biologically plausible spatiotemporal adjustment for training deep spiking neural networks. *Patterns*, 3(6), 2022. 1
- [39] Guobin Shen, Dongcheng Zhao, Yiting Dong, Yang Li, Jin-dong Li, Kang Sun, and Yi Zeng. Astrocyte-enabled advancements in spiking neural networks for large language modeling. arXiv preprint arXiv:2312.07625, 2023. 1
- [40] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Exploiting high performance spiking neural networks with efficient spiking patterns. arXiv preprint arXiv:2301.12356, 2023. 1, 2, 5
- [41] Guobin Shen, Dongcheng Zhao, Tenglong Li, Jindong Li, and Yi Zeng. Are conventional snns really efficient? a perspective from network quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27538–27547, 2024. 1, 6
- [42] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Exploiting nonlinear dendritic adaptive computation in training deep spiking neural networks. *Neural Networks*, 170:190–201, 2024. 5
- [43] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings* of the IEEE/CVF International Conference on Computer Vision, pages 6555–6565, 2023. 7, 8
- [44] Yuchen Wang, Malu Zhang, Yi Chen, and Hong Qu. Signed neuron with memory: Towards simple, accurate and highefficient ann-snn conversion. In *International Joint Confer*ences on Artificial Intelligence, 2022. 6
- [45] Bernard Widrow and István Kollár. Quantization noise: roundoff error in digital computation, signal processing, control, and communications. Cambridge University Press, 2008. 1
- [46] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018. 1, 2, 4
- [47] Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Di He, and Zhouchen Lin. Online training through time for spiking neural networks. Advances in neural information processing systems, 35:20717–20730, 2022. 2, 6, 7
- [48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 34: 12077–12090, 2021. 8
- [49] Xingrun Xing, Boyan Gao, Zheng Zhang, David A Clifton, Shitao Xiao, Li Du, Guoqi Li, and Jiajun Zhang. Spikellm: Scaling up spiking neural network to large language models via saliency-based spiking. arXiv preprint arXiv:2407.04752, 2024. 1
- [50] Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. arXiv preprint arXiv:2307.01694, 2023. 7
- [51] Man Yao, JiaKui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, XU Bo, and Guoqi Li. Spike-driven

- transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 7, 8
- [52] Xingting Yao, Fanrong Li, Zitao Mo, and Jian Cheng. Glif: A unified gated leaky integrate-and-fire neuron for spiking neural networks. Advances in Neural Information Processing Systems, 35:32160–32171, 2022. 5
- [53] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* preprint arXiv:2203.03605, 2022. 6, 8
- [54] Youhui Zhang, Peng Qu, Yu Ji, Weihao Zhang, Guangrong Gao, Guanrui Wang, Sen Song, Guoqi Li, Wenguang Chen, Weimin Zheng, et al. A system hierarchy for brain-inspired computing. *Nature*, 586(7829):378–384, 2020. 1
- [55] Dongcheng Zhao, Guobin Shen, Yiting Dong, Yang Li, and Yi Zeng. Improving stability and performance of spiking neural networks through enhancing temporal consistency. *Pattern Recognition*, 159:111094, 2025.
- [56] Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI conference on artificial* intelligence, pages 11062–11070, 2021. 6, 7
- [57] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on* computer vision and pattern recognition, pages 633–641, 2017. 5, 8, 3
- [58] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. arXiv preprint arXiv:2209.15425, 2022. 1, 2, 3
- [59] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representa*tions, 2023. 5, 7

SpikePack: Enhanced Information Flow in Spiking Neural Networks with High Hardware Compatibility

Supplementary Material

A. Mutual Information Analysis of SpikePack

In this appendix, we provide a formal analysis of the mutual information properties of the proposed *SpikePack* neuron model, comparing it with the traditional Leaky Integrate-and-Fire (LIF) neuron model. This analysis aims to show that *SpikePack* neurons retain more information between pre-synaptic inputs and post-synaptic outputs, thereby reducing information loss during spike transmission.

A.1. Problem Statement

Consider a spiking neuron receiving binary input spikes over T time steps from N pre-synaptic neurons. Let $\mathbf{S}^l \in \{0,1\}^{N \times T}$ denote the input spike matrix, where each element $s_{n,t}^l$ represents the spike from the n-th neuron at time step t. Each spike $s_{n,t}^l$ is assumed to be an independent Bernoulli random variable with parameter p, i.e., $s_{n,t}^l \sim \text{Bernoulli}(p)$. The synaptic weights are represented by $\mathbf{w} \in \mathbb{R}^N$, where each weight w_n is drawn independently from a Gaussian distribution $\mathcal{N}(0,\sigma^2)$.

Our objective is to compute and compare the mutual information $I(\mathbf{S}^l;\mathbf{s}^l)$ between input and output spikes for both SpikePack and LIF neurons.

A.2. Mutual Information in SpikePack Neurons

Accumulated Membrane Potential In the SpikePack neuron, the accumulated membrane potential v_g^l is defined as:

$$v_g^l = \mathbf{w}^\top \mathbf{S}^{l-1} \mathbf{q},\tag{11}$$

where $\mathbf{q}=[\tau^{T-1},\tau^{T-2},\dots,\tau^0]^{\top}$ incorporates the effect of leakage across time steps.

Distribution of v_g^l Given that the input spikes are independent Bernoulli random variables and the weights are independent Gaussian random variables, the accumulated membrane potential v_g^l is a sum of independent random variables. By the Central Limit Theorem, v_g^l approximates a Gaussian distribution when N is large.

Mean of v_q^l :

$$\mu_{v_g^l} = \mathbb{E}[v_g^l] = \sum_{n=1}^N \mathbb{E}[w_n] \sum_{t=1}^T \mathbb{E}[s_{n,t}^l] q_t = 0, \quad (12)$$

since $\mathbb{E}[w_n] = 0$.

Variance of v_q^l :

$$\sigma_{v_g^l}^2 = \mathbb{E}[v_g^l]^2 = \sigma^2 N p(1-p) \left(\sum_{t=1}^T q_t\right)^2,$$
 (13)

where $q_t = \tau^{t-1}$.

Differential Entropy of v_g^l Since v_g^l is approximately Gaussian with variance $\sigma_{v_d^l}^2$, its differential entropy is:

$$h(v_g^l) = \frac{1}{2} \log_2(2\pi e \sigma_{v_g^l}^2). \tag{14}$$

Conditional Entropy $h(v_g^l|\mathbf{s}^l)$ The *SpikePack* neuron generates output spikes \mathbf{s}^l by quantizing the continuous membrane potential v_g^l with a quantization step size θ . This process introduces quantization noise, as v_g^l is mapped to the nearest discrete level defined by θ . Following the approach in [45], we assume that this quantization noise is uniformly distributed over $\left[-\frac{\theta}{2},\frac{\theta}{2}\right]$. This assumption is valid when the quantization step size θ is relatively small compared to the variance of v_g^l , and the signal v_g^l is approximately Gaussian and sufficiently random.

Given that the quantization noise q is uniformly distributed over $\left[-\frac{\theta}{2},\frac{\theta}{2}\right]$, the probability density function of q is:

$$f(q) = \begin{cases} \frac{1}{\theta} & \text{for } -\frac{\theta}{2} \le q \le \frac{\theta}{2}, \\ 0 & \text{otherwise.} \end{cases}$$
 (15)

The conditional entropy $h(v_g^l|\mathbf{s}^l)$ represents the uncertainty introduced by quantizing v_g^l and is equal to the entropy of the quantization noise q over the interval $\left[-\frac{\theta}{2},\frac{\theta}{2}\right]$. The entropy of a continuous uniform distribution is calculated as:

$$h(v_g^l|\mathbf{s}^l) = \int_{-\theta/2}^{\theta/2} -f(q)\log_2(f(q)) dq.$$
 (16)

Substituting $f(q) = \frac{1}{a}$, we get:

$$h(v_q^l|\mathbf{s}^l) = \log_2(\theta). \tag{17}$$

To refine this result, we apply a correction factor for the entropy of the uniform distribution, considering its variance. For a uniform distribution over $\left[-\frac{\theta}{2},\frac{\theta}{2}\right]$, the variance is $\mathrm{Var}(q) = \frac{\theta^2}{12}$ [45]., so the standard deviation is

 $\frac{\theta}{\sqrt{12}}.$ Thus, the correction term $\log_2(\sqrt{12})$ accounts for the spread of the distribution:

$$h(v_q^l|\mathbf{s}^l) = \log_2(\theta) - \log_2(\sqrt{12}).$$
 (18)

This refined expression for the conditional entropy $h(v_g^l|\mathbf{s}^l)$ accurately reflects the quantization effects within the SpikePack neuron model.

Mutual Information Calculation The mutual information between v_q^l and \mathbf{s}^l is:

$$I(v_g^l; \mathbf{s}^l) = h(v_g^l) - h(v_g^l|\mathbf{s}^l) = \frac{1}{2}\log_2\left(\frac{12\sigma_{v_g^l}^2}{\theta^2}\right).$$
 (19)

Since s^l is a deterministic function of v_a^l , we have:

$$I(\mathbf{S}^l; \mathbf{s}^l) = I(v_q^l; \mathbf{s}^l). \tag{20}$$

Thus, the mutual information for the *SpikePack* neuron is:

$$I_{\rm SP} = \frac{1}{2} \log_2 \left(\frac{12\sigma_{v_g^l}^2}{\theta^2} \right). \tag{21}$$

A.3. Mutual Information in LIF Neurons

Approximated Membrane Potential In the LIF neuron, the recursive membrane potential update complicates a direct calculation of mutual information. We approximate the membrane potential at time t as:

$$v_t' = \mathbf{w}^\top \mathbf{s}_t^l, \tag{22}$$

ignoring temporal dependencies and leakage.

Distribution of v_t^\prime Each v_t^\prime is approximately Gaussian with mean zero and variance:

$$\sigma_{v_1'}^2 = \sigma^2 N p (1 - p).$$
 (23)

Probability of Spiking and Entropy of Output Spikes The probability of an output spike at time t is:

$$P(s'_{\text{out},t} = 1) = Q\left(\frac{\theta}{\sigma_{v'_t}}\right),\tag{24}$$

where $Q(\cdot)$ is the Q-function. Using this probability, the entropy of the output spike at each time step is:

$$H(s'_{\text{out},t}) = -P(s'_{\text{out},t} = 1) \log_2 P(s'_{\text{out},t} = 1) - P(s'_{\text{out},t} = 0) \log_2 P(s'_{\text{out},t} = 0).$$
(25)

Upper Bound on Mutual Information Assuming independence across time steps, the total mutual information is bounded by:

$$I_{\text{LIF}} = I(\mathbf{S}^l; \mathbf{s}^l) \le \sum_{t=1}^T H(s'_{\text{out},t}). \tag{26}$$

A.4. Comparative Analysis and Numerical Estimation

Parameter Settings We use the following parameters for both theoretical and numerical estimation:

- Number of pre-synaptic neurons: N=16
- Number of time steps: T = 16
- Weight variance: $\sigma^2 = 1$
- Input spike probability: p = 0.5
- Membrane time constant: $\tau = 2$

SpikePack Mutual Information Calculation Compute $\sigma^2_{v^1_l}$ using Eq. (13):

$$\sigma_{v_{c}^{l}}^{2} = 4\left(2^{16} - 1\right)^{2}. (27)$$

Substitute $\sigma_{v_q^l}^2$ and $\theta = \frac{6\sigma_{v_g^l}}{2^T}$ into Eq. (21):

$$I_{\rm SP} \approx 15.21 \ {\rm bits.}$$
 (28)

LIF Neuron Mutual Information Calculation For the LIF neuron, $\sigma_{v'_t}^2 = 4$ and $P(s'_{\text{out},t} = 1) = Q(0.5) \approx 0.3085$. Using Eq. (25), each time step contributes approximately $H(s'_{\text{out},t}) \approx 0.881$ bits, leading to:

$$I_{\rm LIF} < 16 \times 0.881 = 14.096 \text{ bits.}$$
 (29)

Comparison and Interpretation The mutual information estimates indicate that:

- SpikePack achieves $I_{\rm SP}\approx 15.21$ bits.
- LIF Neuron achieves $I_{\rm LIF} \le 14.096$ bits.

This demonstrates that *SpikePack* retains more information, validating the theoretical analysis.

A.5. Empirical Validation

To validate our theoretical findings, we conducted Monte Carlo simulations to estimate $I(\mathbf{S}^l;\mathbf{s}^l)$ for both neuron models under various configurations of N and T. The results, depicted in Figure 4, Section 3.2, confirm that SpikePack neurons consistently achieve higher mutual information than LIF neurons across different settings, reinforcing the conclusion that SpikePack effectively reduces information loss during spike transmission.

This analysis shows that the *SpikePack* neuron model achieves higher mutual information between input and output spikes than the LIF neuron model. By aggregating information across time steps before spike generation, *SpikePack* reduces information loss and enhances transmission efficiency, supporting more effective information flow in SNNs.

B. Experimental Details

In this section, we provide a comprehensive description of the datasets, model architectures, and hyperparameter settings used in our experiments. This includes details on both static image and neuromorphic datasets, as well as specific training configurations for each task.

B.1. Datasets

We evaluate *SpikePack* on both static and neuromorphic datasets to assess its performance across a range of visual tasks.

Static Datasets

- ImageNet [4]: A large-scale image dataset containing over one million images categorized into 1,000 classes. This dataset provides diverse and complex visual content, which is crucial for evaluating classification performance on high-resolution images. For ImageNet, we resize images to 224 × 224.
- COCO 2017 [27]: A widely-used benchmark for object detection, containing 118,000 training images and 5,000 validation images with 80 object categories. We use this dataset to test *SpikePack* on object detection tasks.
- ADE20K [57]: A semantic segmentation dataset with over 20,000 training images covering 150 classes.
 ADE20K provides a challenging setup for testing dense pixel-wise prediction tasks, such as segmentation.

Neuromorphic Datasets

- CIFAR10-DVS [21]: A neuromorphic adaptation of CIFAR-10, generated using a Dynamic Vision Sensor (DVS) to capture asynchronous event streams. The dataset consists of 10 classes, matching the original CIFAR-10 categories, with each sample transformed into a sequence of events.
- DVS-Gesture [1]: A dataset designed for gesture recognition, containing hand gestures captured from different individuals under varying lighting conditions. The dataset offers dynamic and complex temporal patterns that challenge spiking models.
- N-Caltech101 [35]: This dataset is a neuromorphic version of the Caltech101 object classification dataset, generated through a DVS camera that records event-based sequences for 101 object categories.

B.2. Hyperparameters and Configuration

For our experiments, we evaluate *SpikePack* in two settings: direct training and ANN-to-SNN conversion.

In the direct training setup, we adhere to the settings used by Zhou et al. [58] for comparability and consistency. For ImageNet datasets, the input resolution is set to

 224×224 , unless otherwise noted in the main text. Neuromorphic datasets are resized to 48×48 to streamline computational costs. Batch size is dynamically adjusted according to the specific model architecture, maximizing memory usage without exceeding 40 GB of GPU memory. We employ native Automatic Mixed Precision (AMP) for all training processes to balance computational efficiency and memory usage. The initial learning rate is set to 0.001, and models are trained for 300 epochs unless otherwise specified. The membrane time constant τ is set to 2 by default, and threshold θ is dynamically adjusted as $\theta=T/2^T$, where T is the number of time steps. This approach progressively reduces the threshold over time, creating finer divisions of the input signal, which improves information transmission over longer sequences.

For the ANN-to-SNN conversion experiments, we first calibrate θ by selecting 10% of the training data. This subset is used to set θ in a way that minimizes the risk of overflow during inference. For evaluation, this threshold θ remains fixed to ensure stable performance across the entire test set. During conversion, θ is allocated independently for each channel, enabling fine-grained control over the activation dynamics and improving the robustness of the converted SNN model.

The computation of Synaptic Operations (SOP) follows the same procedure as Zhou et al. [58], where SOP is defined as $SOP = fr \times FLOPs \times T$. Here, fr represents the firing rate, or the proportion of spikes generated over the total possible activations, allowing for a direct comparison of energy efficiency across models with different firing dynamics and time steps.

For object detection and semantic segmentation tasks, we apply the ANN-to-SNN conversion approach, given the high accuracy already achieved through this method. This setup maintains the accuracy benefits of the ANN models while allowing efficient deployment in SNN form, leveraging the sparsity and reduced computational costs enabled by *SpikePack*.

C. Hardware Experiments

To evaluate the performance of *SpikePack* neurons in comparison to traditional Leaky LIF neurons on hardware, we designed a customized digital processor resembling a neuromorphic chip. This processor processes binary spike inputs and synaptic weights, performing event-driven accumulation of synaptic currents. The architecture comprises three primary components: (1) a spike address encoder, which encodes pre-synaptic input spikes to addresses for retrieving the corresponding synaptic weights, (2) an array of processing elements (PEs) with vectorized multiplex-accumulate logic, and (3) parallel neuron node logic responsible for generating output spikes, as depicted in Figure.8. The customized architecture builds upon and extends the

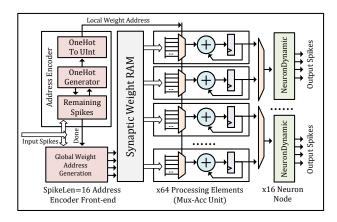


Figure 8. Hardware architecture of the neuromorphic-like processor demo customized for *SpikePack* or LIF neuron.

FireFly-S[24] implementation.

The processor was tailored for both *SpikePack* and LIF neurons, using a shared encoder and PE logic but differing in neuron implementation logic. Table.7 presents the resource consumption of the designs implemented on an XCZU3EG FPGA. In this analysis, we focus on logic resource utilization, excluding on-chip RAM, as synaptic weight data is directly fed from the simulation environment. The device mapping results of two implmentations are shown in Figure.9.

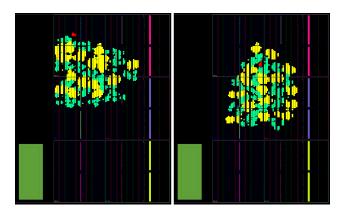


Figure 9. Hardware implementation device map of neuromorphic-like processor for *SpikePack* (right) and LIF (left) neuron on xczu3eg FPGA. Color green area indicates the logic of the processing elements, color yellow indicates the logic of the *SpikePack* or the LIF neuron and color red indicates the logic of the address encoder.

The SpikePack implementation demonstrates a slight reduction in resource consumption compared to the traditional LIF neuron. This efficiency arises from the elimination of the need to store long-term membrane potential in hardware. Additionally, the *SpikePack* implementation consumes less power, operating at 0.808 W compared to 0.816

W for the LIF implementation, both running at 300 MHz.

Table 7. Resource consumption breakdown of customized neuromorphic-like processor for *SpikePack* and LIF neuron.

		LUTs	FFs	CARRY8s
SpikePack	Total	9496	1042	704
	Encode	46	18	0
Spiker ack	PE	4673	1024	256
	Node	4521	0	448
LIF	Total	9850	1302	768
	Encode	46	18	0
	PE	4673	1024	256
	Node	4875	260	512

The ResNet inference latency was measured using a cycle-accurate simulator of the proposed hardware architecture. The spike encoder effectively eliminates redundant spikes, resulting in an inference latency that is strongly correlated with the sparsity level of the spike input. As *SpikePack* inherently produces a more sparse spike output pattern, it achieves lower inference latency and energy per inference compared to the traditional LIF design.