IOE | Intelligent Opto-Electronics

Spiking neural networks for object detection and semantic segmentation across event-driven and frame-based modalities: a review

Anguo Zhang, Hongwei Cao, Na Shan, Jiaqi Wang, Mingbo Pu and Yongduan Song

Citation: Zhang AG, Cao HW, Shan N, et al. Spiking neural networks for object detection and semantic segmentation across event-driven and frame-based modalities: a review. *Intelligent Opto-Electronics* **1**, 250007(2025).

https://doi.org/10.29026/ioe.2025.250007

Received: 11 August 2025; Accepted: 24 September 2025; Published online: 30 September 2025

Related articles

Photonic integrated neuro-synaptic core for convolutional spiking neural network

Shuiying Xiang, Yuechun Shi, Yahui Zhang, Xingxing Guo, Ling Zheng, Yanan Han, Yuna Zhang, Ziwei Song, Dianzhuang Zheng, Tao Zhang, Hailing Wang, Xiaojun Zhu, Xiangfei Chen, Min Qiu, Yichen Shen, Wanhua Zheng, Yue Hao

Opto-Electronic Advances 2023, **6**(11): 230140 doi: 10.29026/oea.2023.230140

Pattern recognition in multi-synaptic photonic spiking neural networks based on a DFB-SA chip

Yanan Han, Shuiying Xiang, Ziwei Song, Shuang Gao, Xingxing Guo, Yahui Zhang, Yuechun Shi, Xiangfei Chen, Yue Hao

Opto-Electronic Science 2023, **2**(9): 230021 doi: 10.29026/oes.2023.230021

All-optical digital logic and neuromorphic computing based on multi-wavelength auxiliary and competition in a single microring resonator

Qiang Zhang, Yingjun Fang, Ning Jiang, Anran Li, Jiahao Qian, Yiqun Zhang, Gang Hu, Kun Qiu

Opto-Electronic Science 2025, (): doi: 10.29026/oes.2025.250003

More related articles in Opto-Electronic Journals Group website

IOE | Intelligent Opto-Electronics
www.oejournal.org/ioe

Review
September 2025, Vol. 1, No. 1

DOI: 10.29026/ioe.2025.250007

Open Access

Spiking neural networks for object detection and semantic segmentation across event-driven and frame-based modalities: a review

Anguo Zhang[1†], Hongwei Cao[2,3†], Na Shan[4,5], Jiaqi Wang[3,6], Mingbo Pu[2,3,6] and Yongduan Song[7]*

Abstract: Spiking neural networks (SNNs), drawing inspiration from the energy-efficient and event-driven processing of biological brains, are emerging as a compelling alternative to traditional artificial neural networks (ANNs) for resource-constrained artificial intelligence (AI) applications. Their intrinsic properties, including low power consumption, ultra-low latency, and native spatio-temporal information processing capabilities, position them as ideal candidates for critical computer vision tasks such as real-time object detection and semantic segmentation, especially at the edge. This review systematically explores the fundamental principles of SNNs, including their unique neuron models and information encoding schemes, contrasting them with the operational paradigms of ANNs. We delve into the sophisticated mathematical formulations underpinning key SNN neuron models and the intricate learning dynamics that differentiate SNNs. A significant portion is dedicated to meticulously dissecting recent architectural innovations in SNNs tailored for image object detection and semantic segmentation. This includes an in-depth analysis of pure SNN convolutional networks, pragmatic hybrid SNN-ANN models, and the cutting-edge integration of attention mechanisms and Transformer-based designs. Furthermore, we provide an enhanced exposition of crucial training algorithms, such as advanced surrogate gradient methods and spiking batch normalization, highlighting their theoretical underpinnings and practical implications. Finally, this review synthesizes the current performance benchmarks, identifies persistent research challenges, and delineates promising future directions, particularly emphasizing the synergistic co-design of SNN algorithms and neuromorphic hardware. We argue that SNNs, while not yet universally outperforming ANNs, hold immense potential to revolutionize AI in dynamic, resource-limited environments, becoming a cornerstone of next-generation intelligent systems.

Keywords: spiking neural networks (SNNs); object detection; semantic segmentation; neuromorphic computing; bio-inspired AI

Citation: Zhang AG, Cao HW, Shan N, et al. Spiking neural networks for object detection and semantic segmentation across event-driven and frame-based modalities: a review. *Intell Opto-Electron* **1**, 250007 (2025).

Introduction

The remarkable advancements in deep learning over the past decade have profoundly reshaped the landscape of artificial intelligence (AI), particularly in computer vision[1−3] and natural language processing[4−7]. Deep artificial neural networks (ANNs) have demonstrated unparalleled capabilities in tasks ranging from image recognition and speech processing to complex decision-making[8−11]. However, this exceptional performance comes at a significant cost: ANNs are inherently resource-intensive, demanding substantial computational power, large datasets, and considerable

Received: 11 August 2025
Accepted: 24 September 2025
Published online: 30 September 2025

[1]Interdisciplinary Institute of Medical Engineering, Fuzhou University, Fuzhou 350108, China; [2]State Key Laboratory of Optical Field Manipulation Science and Technology, Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China; [3]Research Center on Vector Optical Fields, Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China; [4]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430205, China; [5]Wuhan Eternal Technologies Co., Ltd, Wuhan 430074, China; [6]College of Materials Science and Opto-Electronic Technology , University of Chinese Academy of Sciences, Beijing 100049, China; [7]School of Automation, Chongqing University, Chongqing 400030, China.
[†]These authors contributed equally to this work.
*Correspondence: YD Song, E-mail: ydsong@cqu.edu.cn

energy consumption[12–14]. This intrinsic limitation poses a formidable challenge for real-time and edge AI applications, such as autonomous vehicles[15–18], unmanned aerial vehicles (UAVs)[19–22], and collaborative robots[23–27], where energy efficiency and low latency are paramount for operational endurance and responsiveness. The burgeoning demand for edge AI, driven by the need for fast, real-time responses, enhanced data privacy, and reduced power consumption, necessitates novel AI paradigms that can operate efficiently within strict power and computational budgets[28–31]. This inherent limitation of traditional DNNs is not an insurmountable barrier, but rather a powerful catalyst for innovation, fostering the co-development of alternative solutions across hardware, algorithmic, and application layers.

In direct response to these limitations, SNNs have emerged as a highly promising, bio-inspired computing paradigm, often heralded as third-generation neural networks[32,33]. SNNs distinguish themselves by mimicking biological neural networks, employing neuron models that communicate via discrete, asynchronous electrical pulses-spikes-rather than continuous, real-valued activations. This event-driven, sparse computational model inherently offers remarkable energy efficiency and native compatibility with temporal information encoding[34–41]. As a crucial link between neuroscience and machine learning, SNNs portend a future where AI development will draw more deeply from biological intelligence, potentially leading to the emergence of more general, efficient, and robust intelligent systems[42–44].

Within computer vision, object detection and semantic segmentation are foundational tasks, crucial for applications spanning video surveillance, autonomous driving, and medical image analysis[45–48]. Object detection precisely identifies and localizes objects[49–51], while semantic segmentation assigns class labels to every pixel for fine-grained scene understanding[52–54]. However, real-world scenarios introduce significant complexities, including scale variations, illumination changes, and ambiguous boundaries[55]. Traditional DNNs, despite their impressive results, often struggle with these complexities, imposing severe computational demands that render them unsuitable for resource-constrained environments. This context positions SNNs, with their promise of low power consumption and low latency, as a potential transformative solution[56]. The demand for real-time, dynamic scene understanding aligns exceptionally well with event cameras, which provide high temporal resolution and event-driven data streams[57–61]. SNNs, as an event-driven computing paradigm, exhibit a natural synergy with such data[62–64], suggesting superior efficiency for dynamic semantic segmentation compared to frame-based ANNs.

Historically, SNNs have lagged behind ANNs in complex tasks, but recent breakthroughs have dramatically narrowed this performance gap, with SNNs achieving comparable or superior results on specific benchmarks while significantly reducing energy consumption[65–68]. This rapid improvement is driven by convergent advances in neuromorphic hardware[69], event camera technology[70], and sophisticated algorithmic developments in SNN architectures and learning strategies. The diminishing performance disparity signifies a pivotal shift in SNN technology, from theoretical exploration to practical viability, foreshadowing a surge in SNN research geared towards practical deployment and real-world applications. Given this evolving landscape, this comprehensive review is designed to provide an in-depth exploration of the latest advancements in SNNs for image object detection and semantic segmentation. It systematically covers fundamental SNN concepts, including neuron models and information encoding schemes, and elucidates their foundational differences from ANNs. We then dissect recent architectural innovations, advanced learning strategies, their synergistic relationship with neuromorphic hardware, quantitative performance benchmarks, and critical future research directions.

Specifically, this review makes the following key contributions: 1) a systematic and critical analysis of the latest SNN architectures for object detection and semantic segmentation, with a special focus on hybrid and Transformer-based models, 2) an in-depth discussion of advanced training algorithms tailored for SNNs, such as surrogate gradients and spiking batch normalization, and 3) a unique focus on the synergy between SNNs and optoelectronic sensors (e.g., event cameras), aligning with the cutting-edge field of Intelligent Opto-Electronics.

Spiking neural network fundamentals

SNNs draw profound inspiration from the intricate architecture and dynamic functionality of biological neural systems[71], distinguishing themselves from traditional ANNs by employing neuron models that communicate via discrete, asynchronous electrical pulses, known as “spikes” or “action potentials”[33]. This fundamental paradigm shift allows SNNs to process information in a unique manner, leveraging the precise timing and frequency of these spikes to encode and transmit information effectively.

Core concepts of spiking neuron dynamics

The operational paradigm of SNNs is deeply rooted in neurobiological principles, offering a more biologically plausible model of computation. Central to SNNs is the concept of a neuron’s internal state, governed by its membrane potential ($V_\mathrm{m}(t)$), which accumulates incoming synaptic currents over time. When this potential reaches or exceeds a predefined threshold (V_th), the neuron fires a discrete spike, after which its membrane potential is typically reset. This spike generation process introduces a crucial non-linearity to the neuron’s dynamics. The general form for membrane potential dynamics can be described by a first-order differential equation:

$$\tau_{\mathrm{m}}\frac{\mathrm{d}V_{\mathrm{m}}\left(t\right)}{\mathrm{d}t}=-\left(V_{\mathrm{m}}\left(t\right)-V_{\mathrm{rest}}\right)+R_{\mathrm{m}}I_{\mathrm{syn}}\left(t\right)\ , \qquad (1)$$

where τ_{m} is the membrane time constant, V_{rest} is the resting membrane potential, R_{m} is the membrane resistance, and $I_{\mathrm{syn}}\left(t\right)$ is the total synaptic current arriving at the neuron. The strength of connections between neurons is modulated by synaptic weights, which govern the influence of an incoming spike on the postsynaptic neuron's membrane potential. Synaptic inputs can be excitatory or inhibitory. When a presynaptic neuron fires a spike, it induces a transient change in the membrane potential of the postsynaptic neuron, known as the post-synaptic potential (PSP). The ability of synapses to strengthen or weaken over time is termed synaptic plasticity, which forms the biological basis of learning and memory[72].

The inherent sparsity and event-driven nature of SNNs, where computations occur only upon spike firing, directly contribute to their exceptional energy efficiency[73]. This stands in stark contrast to ANNs, which typically involve continuous activation and dense computations across all neurons at every timestep. However, the discrete and complex spike dynamics of SNNs pose significant challenges for direct application of traditional gradient-based learning algorithms like backpropagation, necessitating the development of specialized training strategies such as surrogate gradients.

Key spiking neuron models

Diverse spike-based neuron models exist, each offering varying degrees of biological realism, computational complexity, and efficiency, serving as the fundamental building blocks of SNN architectures[33].

Integrate-and-fire (IF) and leaky integrate-and-fire (LIF) models

The IF model is the simplest conceptualization of a spiking neuron. It functions as a perfect integrator, where its membrane potential $V_{\mathrm{m}}\left(t\right)$ accumulates incoming synaptic current over time. Once $V_{\mathrm{m}}\left(t\right)$ reaches a predefined threshold V_{th}, the neuron fires a spike and its potential is reset. However, the basic IF model lacks a crucial biological feature: the passive decay of membrane potential over time[74].

The LIF model addresses this by incorporating a "leakage" term, making it both more biologically realistic and the most widely adopted neuron model in SNN research due to its excellent balance of plausibility and computational efficiency[75]. The LIF model simulates how a neuron integrates incoming current while its membrane potential simultaneously "leaks" back towards a resting state. The dynamics of the LIF neuron's membrane potential $V_{\mathrm{m}}\left(t\right)$ are described by the differential equation:

$$\tau_{\mathrm{m}}\frac{\mathrm{d}V_{\mathrm{m}}\left(t\right)}{\mathrm{d}t}=-\left(V_{\mathrm{m}}\left(t\right)-V_{\mathrm{rest}}\right)+I_{\mathrm{in}}\left(t\right)\ , \qquad (2)$$

where the term $-\left(V_{\mathrm{m}}\left(t\right)-V_{\mathrm{rest}}\right)$ represents the passive leak, causing the potential to exponentially decay towards the resting potential V_{rest} with a membrane time constant τ_{m}. $I_{\mathrm{in}}\left(t\right)$ is the total input current from presynaptic neurons. Upon spiking at time t_{f} (when $V_{\mathrm{m}}\left(t_{\mathrm{f}}\right)\geq V_{\mathrm{th}}$), the neuron fires, and its membrane potential is instantly reset to a lower value V_{reset} (where $V_{\mathrm{reset}}\leq V_{\mathrm{rest}}$) and is often clamped for a brief refractory period to prevent immediate re-firing.

While the IF and LIF models' simplicity is advantageous, they face limitations in complex deep networks, primarily because their fixed parameters (e.g., τ_{m}, V_{th}) may require extensive manual tuning and can limit their expressive power and adaptability. Nevertheless, their computational efficiency has made them a cornerstone in many pioneering SNN architectures, such as Spiking-YOLO[76] and Spiking-YOLOX[77] for object detection. They are also used in Spiking CenterNet[78] for event data and SpikeFPN[79] for adaptive threshold mechanisms. For classification and segmentation, LIF neurons are employed in Spiking-SSeg-Net[80] and Spiking-UNet variants for image segmentation[81,82].

Parameterized leaky-integrate-and-fire (PLIF) model

The PLIF neuron, a further extension of the LIF model, introduces learnable parameters for the membrane time constant (τ_{m}) and the threshold potential (V_{th}) that can be optimized during training. This allows the neuron's dynamics to adapt more flexibly to the data and task requirements, improving the model's expressive power and training stability[83]. PLIF models have been used in embedded SNN object detection[84] and in LT-SNN[85], which optimizes learnable thresholds online. The PLIF model is also a key component in EvSegSNN[86] for semantic segmentation. The PLIF model's learnable parameters enable it to capture more complex neuronal behaviors without significantly increasing the computational overhead, offering a bridge between the simplicity of LIF and the adaptability of more complex models.

Bistable integrate-and-fire (BIF) model

The BIF neurons, introduced by Yasir et al.[87], represent a novel approach by exhibiting two stable states, which enhances information transmission and stability. This mechanism improves temporal coding and significantly enhances detection performance by optimizing spike utilization and encoding more information per spike.

Multi-threshold spiking neuron

Multi-threshold spiking neurons, as introduced in Lei et al.[88] and Li et al.[89], fire multiple spikes based on a series of predefined thresholds. This mechanism is designed to enhance information transmission in SNNs, especially within complex architectures like Mask2Former. By allowing neurons to fire upon crossing different thresholds, the model's output can better align with ANN activations, thereby streamlining conversion and training processes. This approach, when combined with connection-wise normalization, helps prevent inconsistent firing rates in skip connections, ensuring faithful information representation across the network.

Dynamic threshold leaky-integrate-and-fire (DT-LIF) model

To enhance adaptability beyond fixed thresholds, the DT-LIF neuron dynamically adjusts its firing threshold based on past activity, mimicking biological neuronal adaptation and firing-rate homeostasis[90]. The dynamic threshold $V_{\mathrm{th}}\,(t)$ for a DT-LIF neuron can be modeled as

$$V_{\mathrm{th}}\,(t)=V_{\mathrm{th},0}+\Sigma_{\mathrm{k}}\eta_{\mathrm{k}}\mathrm{e}^{-(t-t_{\mathrm{k}})/\tau_{\mathrm{adapt}}}\,,\tag{3}$$

where $V_{\mathrm{th},0}$ is the baseline threshold, η_{k} is an increment added to the threshold after each spike fired at time t_{k}, and τ_{adapt} is the adaptation time constant governing the exponential decay of the threshold. This dynamic adjustment significantly enhances inference speed and accuracy by preventing neurons from firing excessively or becoming "dead" due to consistently high or low membrane potentials. After each spike, the threshold transiently increases and then gradually decays. The DT-LIF model represents a balance: it is "bio-inspired" yet incorporates engineering adjustments for computational efficiency and training performance, reflecting a core theme in SNN research to compromise "pure biological realism" for "computational feasibility". This model is employed in DT-LIF Based SSD[68] to improve detection accuracy and inference speed.

Analog spiking neuron

In Ma et al.'s[91] analog Spiking U-Net, an analog spiking neuron is proposed which modifies the firing positions of neurons and transfers information in floating-point signals, aiming to preserve detailed information. This model integrates analog CBAM (convolutional block attention module) and spiking ViTBlock (vision transformer block) to enhance semantic segmentation. The analog CBAM is specifically designed to handle floating-point signals from ANNs before conversion to spikes, enabling the use of traditional ANN modules without corrupting spike distribution. This innovative approach seeks to bridge the information gap often encountered when converting continuous ANN activations to discrete SNN spikes.

The continuous evolution of SNN neuron models directly addresses inherent training difficulties and performance limitations[92]. The progression from basic LIF to adaptable PLIF, and the introduction of multi-threshold, analog, and NSNP neurons, reflect ongoing efforts to balance computational efficiency with complex network performance, bridging the gap with ANNs and enhancing task-specific capabilities.

Information encoding schemes

In SNNs, information is represented and communicated through spikes, necessitating efficient encoding schemes to translate input data into spike trains and decode output spike trains into meaningful representations[93]. These schemes leverage SNNs' unique spatio-temporal properties, typical rate encoding and temporal encoding mechanisms are as shown in Fig. 1.

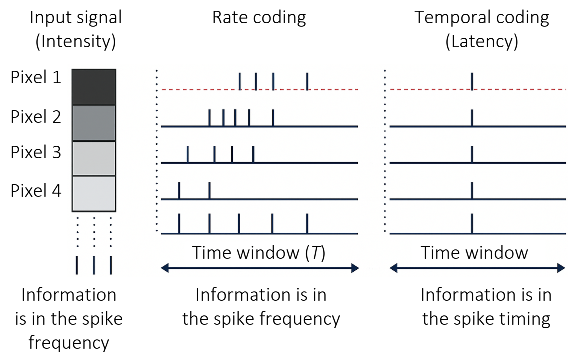
Input signal (Intensity)
Rate coding
Temporal coding (Latency)
Pixel 1
Pixel 2
Pixel 3
Pixel 4
Time window (T)
Time window
Information is in the spike frequency
Information is in the spike frequency
Information is in the spike timing

Fig. 1 | The rate encoding and temporal encoding mechanisms of input signal for SNNs.

Rate encoding

Rate encoding schemes represent information through the average firing rate or frequency of spikes within a given time window, where a higher firing rate typically signifies a stronger signal. The firing rate R of a neuron can be simply expressed as

$$R=\frac{N_{\mathrm{spikes}}}{T_{\mathrm{window}}}\,,\tag{4}$$

where N_{spikes} is the number of spikes fired within a time window T_{window}. While simpler to implement and often used for compatibility with rate-based ANN concepts, this method can potentially lose the fine-grained temporal information inherent in spike sequences[94]. Rate-based spike coding is used in SpikiLi[95] for LiDAR-based 3D object detection. However, rate encoding faces challenges with datasets of varying intensities[81]. To address this, constant current injection[82] or normalized voxel grids[96] are used to ensure uniformity in scale and mitigate outliers.

Temporal encoding

Temporal encoding schemes leverage the precise timing of spikes or the relative timing between spikes to represent information. This approach is highly information-rich and inherently compatible with the dynamic nature of biological neural networks. This includes latency coding, where information is encoded in the time of the first spike relative to a reference point or stimulus (shorter latencies often correspond to stronger input signals), and rank-order coding, which encodes information in the relative order of firing of different neurons (the first neuron to fire might carry the most salient information). Temporal encoding is particularly powerful for capturing dynamic spatio-temporal information, making SNNs naturally adept at processing time-series data and dynamic scenes. Its energy efficiency stems from the fact that neurons only fire when necessary, minimizing activity. Spike-time-dependent integrated (STDI) coding, proposed by Qu et al.[97], further augments information capacity in individual spikes for ultralow-latency SNNs.

The choice between encoding schemes involves a trade-off between performance and energy efficiency. Rate encoding is simpler but may sacrifice temporal precision, whereas temporal encoding is information-rich but can be more complex to implement and train. This trade-off directly influences model accuracy and computational efficiency, driving future research to explore more efficient and task-adaptive hybrid encoding schemes[98].

Fundamental differences and advantages over ANNs

SNNs fundamentally diverge from ANNs in several critical aspects, which collectively contribute to their unique advantages and challenges.

Fundamental differences

SNNs communicate via discrete, binary, and asynchronous electrical pulses (spikes), contrasted with ANNs' continuous, real-valued, and typically synchronous activations. This event-driven nature means SNNs process information only when a spike occurs, leading to sparse and asynchronous computations, unlike ANNs' dense, synchronous processing. SNNs are designed to more closely mimic biological brains, guiding their core architectural and learning principles, while ANNs are abstract mathematical models. The discrete and non-differentiable nature of spike generation in SNNs complicates direct application of traditional gradient-based learning algorithms like backpropagation, necessitating specialized SNN training techniques.

Key advantages

As shown in Fig. 2, energy efficiency is a cornerstone advantage, as SNNs transmit information and perform computations only when a neuron fires, leading to significantly fewer operations and orders of magnitude lower energy consumption, crucial for power-constrained edge AI devices[99−100]. For instance, sparse compressed SNN accelerators have achieved 26x model size reduction and high energy efficiency for object detection[101]. Spiking-YOLO adaptations have shown 280x less energy consumption on TrueNorth[102]. Spiking-UNet achieved a 10x energy reduction compared to its CNN counterpart[82]. SNNs replace high-power MAC operations with more energy-efficient AC operations, particularly for neuromorphic hardware[91]. PSSD-Transformer consumes 17.76× less energy than ANN-based models[103].

Low latency is another key benefit, on specialized neuromorphic hardware, SNNs can achieve extremely low latency processing, outperforming ANNs in real-time applications by processing information as events occur rather than waiting for full-frame data. SUHD[97], an ultralow-latency and high-accuracy SNN for object detection, achieves 750× timestep reduction and 30% mAP enhancement. Low latency in SNNs is also achieved by using Neurons-Shared Blocks and transfer learning, enabling rapid inference with fewer time steps[80].

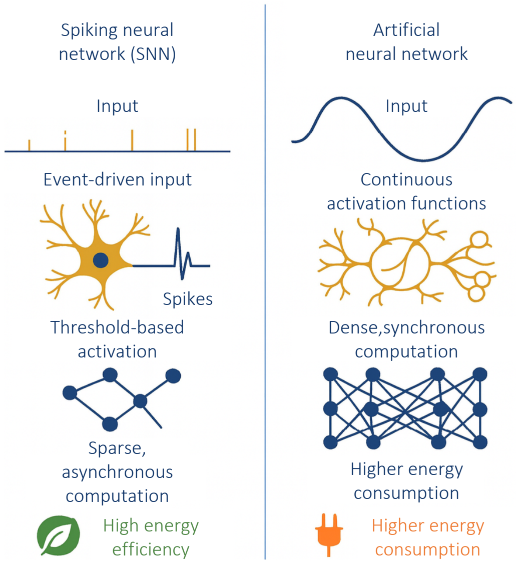
Spiking neural network (SNN)
Input
Event-driven input
Spikes
Threshold-based activation
Sparse, asynchronous computation
High energy efficiency
Artificial neural network
Input
Continuous activation functions
Dense,synchronous computation
Higher energy consumption
Higher energy consumption

Fig. 2 | Difference and advantages of SNN compared to ANN.

Sparsity in SNNs is a natural characteristic inherited from biological neural networks, contributing to their computational efficiency and reduced memory footprint. Addressing information loss in sparse spiking is a key area of research, with solutions like spike-driven deformable transformer encoder (SDTE) and spike-driven mask embedding (SDME) enhancing segmentation performance[88]. Pure sparse self attention (PSSA) and dynamic spiking membrane shortcut (DSMS) ensure spike-based processing without floating-point computations[103].

Temporal data processing is an intrinsic strength of SNNs, as they encode information in spike timings, making them well-suited for processing spatio-temporal data and complex dynamic patterns, ideal for sequential or event-based sensory inputs. SNNs for image segmentation demonstrate dynamic event-driven processing and temporal axis capacity, opening new horizons for models with exponential memorization[81]. Spiking-LSTM models combine SNN and LSTM to capture spatio-temporal information effectively for tasks like hyperspectral image segmentation[104].

Compatibility with event-driven sensors is another key advantage. SNNs' event-driven operational paradigm makes them highly compatible with data streams from event cameras, which generate asynchronous event-based data, offering superior performance in challenging conditions like high-speed motion, high dynamic range, and low light[64,105]. EvSegSNN[86] highlighted SNNs' suitability for event-based sensors due to their asynchronous spike computation and speed of spread, making them ideal for low-power, real-time semantic segmentation tasks. Beyond image-based tasks, SNNs also process raw sensor data directly, as exemplified by[106] for automotive radar object detection.

The spiking object detection and semantic segmentation pipelines with event-driven data input is presented in Fig. 3, a comparative overview of SNNs and ANNs/DNNs, highlighting their distinct characteristics and advantages, is presented in Table 1.

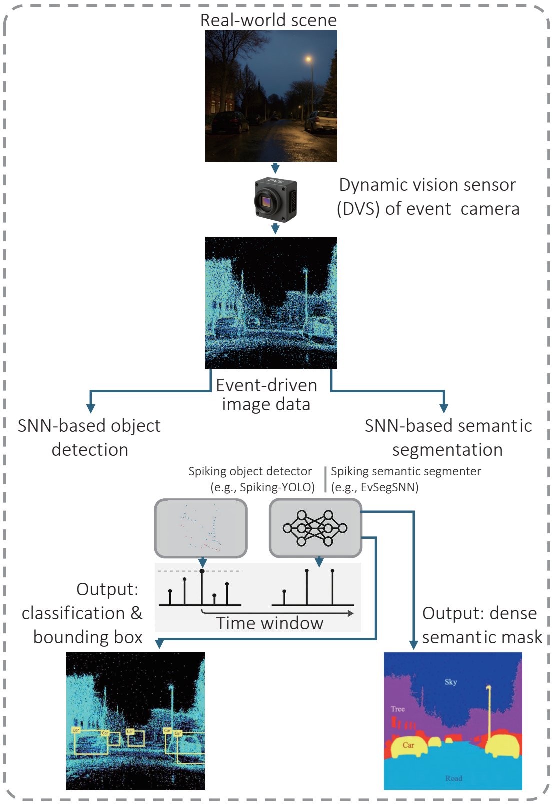
Real-world scene
Dynamic vision sensor (DVS) of event camera
Event-driven image data
SNN-based object detection
SNN-based semantic segmentation
Spiking object detector (e.g., Spiking-YOLO)
Spiking semantic segmenter (e.g., EvSegSNN)
Output: classification & bounding box
Time window
Output: dense semantic mask
Sky
Tree
Car
Road

Fig. 3 | The spiking object detection and semantic segmentation pipelines with event-driven data input. This concrete visual representation illustrates the entire process, from a real-world scene captured by an event camera (generating sparse, point-like events) to the processed outputs (bounding boxes for object detection and pixel-wise masks for semantic segmentation).

While the biological plausibility that garners SNNs much praise is simultaneously the root cause of their primary challenges, training difficulty and non-differentiability due to discrete spiking and complex dynamics, this fundamental trade-off lies at the heart of SNN research[107,108]. The discrete nature and complex dynamics of SNN spikes render traditional gradient descent methods impractical, propelling researchers to develop alternative gradient-based or gradient-free approaches. These methods, by necessity, often involve some level of compromise on pure biological realism in favor of computational feasibility. Crucially, the full realization of SNNs' energy efficiency and low-latency advantages is heavily contingent on specialized neuromorphic hardware[109,110]. Traditional computing platforms like GPUs and CPUs are not optimized for event-driven, sparse computation, thus hindering SNNs from demonstrating their full potential energy savings. Consequently, the future trajectory of SNN development is inextricably linked to the synergistic co-design of algorithms and hardware, alongside the broader commercialization and accessibility of neuromorphic chips[111].

SNN architectures for object detection

Object detection, a cornerstone task in computer vision, demands both high accuracy and real-time performance. SNNs, with their inherent energy efficiency and event-driven nature, are uniquely positioned to address the computational demands of deploying object detection models on edge devices. This section reviews the evolution of SNN architectures for object detection, from early conceptual models to advanced hybrid and Transformer-based designs.

Early SNN architectures for vision tasks

The initial foray of SNNs into computer vision primarily focused on simpler recognition and classification tasks, such as handwritten digit recognition (e.g., MNIST dataset) or basic pattern classification[112–113]. Meftah et al.[114] explored SNNs for image segmentation and edge detection using

Table 1 | Comparison of SNNs and ANNs/DNNs for object detection and semantic segmentation.

Feature	SNNs	ANNs/DNNs
Communication mechanism	Discrete spikes, asynchronous	Continuous values, synchronous
Neuron activation	Threshold-based, event-driven	Continuous activation functions (e.g., ReLU)
Information processing	Sparse, temporal (Timing/Frequency)	Dense, rate-based (Activation Strength)
Learning paradigm	STDP, surrogate gradients, ANN-to-SNN conversion	Backpropagation (Gradient Descent)
Energy consumption	Low (inherently efficient)	High (resource-intensive)
Latency	Ultra-low (especially on neuromorphic hardware)	Higher (due to synchronous processing)
Biological Plausibility	High (mimics biological brains)	Low (abstract mathematical models)
Training difficulty	High (non-differentiable spikes)	Lower (well-established methods)
Hardware compatibility	Neuromorphic processors, edge devices	GPUs, CPUs, TPUs
Typical performance (accuracy)	Rapidly improving, approaching/exceeding ANNs on specific tasks	High, state-of-the-art across many tasks
Data modality suitability	Event-based data, time-series	Frame-based data, Static images

unsupervised Hebbian-based winner-take-all learning with LIF neurons. Pioneering conversion work by Cao et al.[115] demonstrated an early method to convert deep CNNs into SNNs for energy-efficient object recognition, achieving two orders of magnitude lower energy consumption compared to FPGA-based CNNs while maintaining similar accuracy. These foundational efforts were crucial for demonstrating the feasibility of spike-based computation in the visual domain, setting the groundwork for adapting successful deep learning concepts to the SNN framework despite the inherent architectural simplicity that limited their applicability to more complex object detection problems.

Pure SNN convolutional networks (S-CNNs)

As depicted in Fig. 4, inspired by the immense success of Convolutional Neural Networks (CNNs) in conventional computer vision, researchers began translating these powerful architectures into the SNN paradigm, giving rise to deep Spiking Convolutional Neural Networks (S-CNNs). These models have demonstrated notable energy efficiency advantages, particularly in event-driven object detection tasks. The development trajectory of S-CNNs involved adapting or simplifying successful ANN detectors, leading to a strategic shift within the SNN field toward optimizing for its unique characteristics.

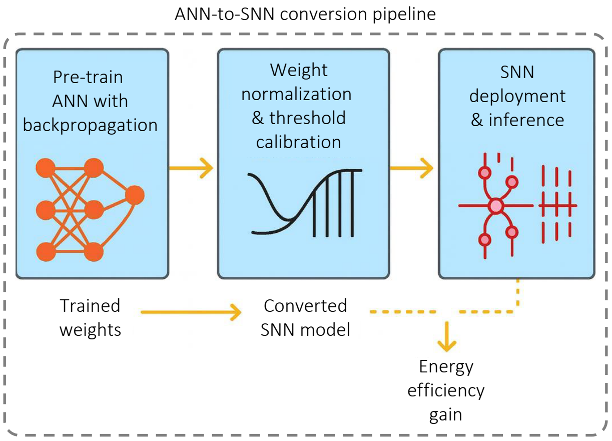
ANN-to-SNN conversion pipeline
Pre-train ANN with backpropagation
Weight normalization & threshold calibration
SNN deployment & inference
Trained weights
Converted SNN model
Energy efficiency gain

Fig. 4 | Generic ANN-to-SNN conversion pipeline. This diagram illustrates the transformation process, highlighting how a pre-trained ANN is adapted to an SNN, typically involving threshold mapping and weight transfer to leverage existing deep learning successes.

Spiking-YOLO and its variants

Spiking-YOLO[76] is an SNN adaptation of the popular YOLO (You only look once) object detection architecture. It introduces a “meta-SNN block,” a channel normalization scheme, and unbalanced threshold sign neurons. The primary motivation was to leverage YOLO’s efficiency and accuracy for object detection within the SNN framework, addressing the inefficiencies of traditional normalization methods and enhancing adaptability to diverse datasets. This aimed to resolve challenges in building deep SNNs capable of complex vision tasks while maintaining the energy efficiency inherent to spiking neurons. It simplifies the YOLO architecture to suit SNN characteristics by incorporating specific SNN-friendly components like the meta-SNN block and custom normalization. This allows for direct training of deep SNNs without ANN-to-SNN conversion. However, it still faces high training complexity and an accuracy gap compared to its ANN counterparts. Building on this, Bi et al.[116] adapted YOLOv5’s C3 module using LIF neurons, achieving 2.47× lower power consumption and improved accuracy for foreign object detection on overhead power lines. Liu et al.[102] further adapted spiking-YOLO for mobile robot deployment, demonstrating 280× less energy consumption on TrueNorth. Miao et al.[77] advanced spiking-YOLOX by integrating ternary signed spiking neurons and fast fourier convolution (FFC) for enhanced feature extraction and state-of-the-art object detection. Additionally, Qu et al.[97] proposed SUHD, an ultralow-latency SNN that enhances SPPF conversion efficiency through timestep compression and spike-time-dependent integrated (STDI) coding.

Multi-scale spiking detectors

The multi-scale spiking detector (MSD) framework[117] integrates spiking multi-scale fusion with dedicated spiking detectors. Traditional SNNs often struggle with detecting objects of varying sizes in complex scenes due to their focus on local features. MSD was developed to enhance deep feature extraction across multiple scales, which is crucial for robust object detection. It achieves high performance with low energy consumption by utilizing On-chip neuromorphic network blocks (ONNB) and a multi-scale spiking fusion mechanism, directly training deep SNNs. This allows for a more comprehensive understanding of visual scenes. As a related model, the spiking fusion object detector (SFOD)[118] combines a spiking denseNet backbone with an SSD (single shot MultiBox detector) detection head. Its innovative spiking fusion module enables multi-scale feature fusion not only spatially but also temporally, improving detection accuracy for dynamic objects by fusing transient movements from shallow layers with broader actions from deeper layers. Fan et al.[119] introduced SFDNet, a fully spiking RGB-event fusion-based detection network featuring the leaky integrate-and-multi-fire (LIMF) neuron model and a multi-scale hierarchical spiking residual attention network, achieving state-of-the-art low-power and robust detection. Furthermore, Fan et al.[120] introduced SpikeDet, which optimizes firing patterns using a Membrane-based deformed shortcut residual network (MDSNet) and spiking multi-direction fusion module (SMFM), achieving high AP with reduced power consumption.

Other notable contributions to pure SNNs for object detection

Pure SNNs have seen diverse advancements for object detection. Bulzomi et al.[121] proposed a lightweight SNN using visual attention mechanisms to filter noise, achieving a 24× smaller model size. Courtois et al.[84] demonstrated embedded SNN object detection with a SpikeThin-VGG backbone on an FPGA-based SPLEAT accelerator, achieving 490 mJ/prediction for automotive event data. Bodden et al.[78] introduced Spiking CenterNet, utilizing an M2U-Net-based decoder and knowledge distillation to achieve 2.6% higher mAP. Zhang et al.[79] proposed SpikeFPN for automotive event-based object detection, which uses an FPN architecture and an adaptive threshold mechanism, achieving 0.477 mAP on the GEN1 dataset. Mohapatra et al.[95] presented SpikiLi, an SNN for LiDAR-based 3D object detection, leveraging CNN-to-SNN conversion and quantized weights for efficient autonomous driving. Feng et al.[122] proposed a multi-patch localization SNN for infrared drone object detection, decoupling classification and localization tasks to achieve 98.9% accuracy with low power. Lien and Chang[101] demonstrated a sparse compressed SNN accelerator, achieving 26× model size reduction and 1.05 mJ/frame energy efficiency. Su et al.[123] proposed EMS-YOLO, a deep directly-trained SNN for object detection that achieves ANN-comparable performance with 5.83x less energy. Zhang et al.[124] introduced SG ResNet with a binary selection gate, addressing gradient vanishing and achieving high accuracy. Qu et al.[97] proposed SUHD, an ultralow-latency SNN achieving 750× timestep reduction and 30% mAP enhancement. Yasir et al.[87] introduced BN-SNN, integrating bistable integrate-and-fire (BIF) neurons to enhance information transmission and improve detection performance. Wang et al.[125] developed Spike-BRGNet, an event-based semantic segmentation network for traffic scenes, featuring a three-branch spiking encoder and a spiking multi-scale context aggregation (SMSCA) module, outperforming SOTA SNN methods by +1.57%-1.91% mIoU while consuming 17.76× less energy than ANN-based models.

A notable limitation of early spiking CNNs is their propensity to focus on local and single-scale features. This inherent bias makes it challenging to achieve high detection accuracy, especially for objects with varying sizes or in complex, cluttered scenes. While traditional CNNs excel at local feature extraction, their SNN counterparts inherited this characteristic, which proves particularly problematic for object detection where understanding the global scene context is paramount for accurate localization and classification. This limitation in feature representation has directly spurred the development of more advanced architectures, including the integration of feature pyramids and Transformer-based designs. The early development trajectory of S-CNNs involved adapting or simplifying successful ANN detectors (e.g., YOLO, SSD). While this expedited SNN application to complex tasks, its inherent limitations drove researchers to explore more “native” SNN designs like MSD’s spike-based multi-scale fusion and ARSNN’s unique approach to temporal alignment loss. This indicates a strategic shift within the SNN field toward optimizing for its unique characteristics rather than mere transplantation.

Hybrid SNN-ANN Architectures

Recognizing the complementary strengths of SNNs (energy efficiency, temporal processing) and ANNs (high accuracy, robust training, mature learning algorithms), researchers have explored integrating these two paradigms into hybrid architectures (typically as shown in Fig. 5). The core philosophy behind these hybrid models is to leverage SNNs’ power-efficient, event-driven processing capabilities, typically for extracting low-level, spatio-temporal features from event data, with ANNs’ established efficient learning and powerful representation capabilities, often for high-level tasks like object classification and bounding box regression. Hybrid architectures represent a pragmatic engineering compromise, meticulously crafted to mitigate the performance disparity of SNNs while concurrently preserving their inherent efficiency advantages.

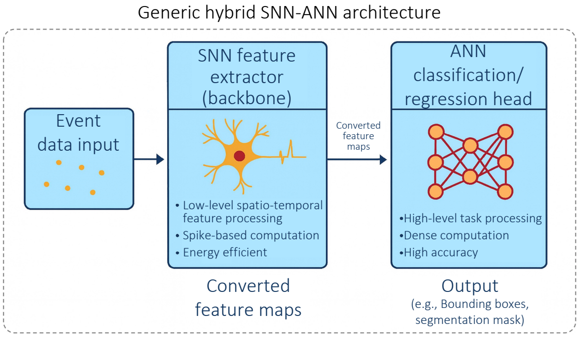
Generic hybrid SNN-ANN architecture
Event data input
SNN feature extractor (backbone)
• Low-level spatio-temporal feature processing
• Spike-based computation
• Energy efficient
Converted feature maps
Converted feature maps
ANN classification/ regression head
•High-level task processing
•Dense computation
•High accuracy
Output
(e.g., Bounding boxes, segmentation mask)

Fig. 5 | Generic hybrid SNN-ANN architecture. This conceptual diagram illustrates how SNN components, often used for efficient low-level feature extraction, are integrated with ANN components, typically handling high-level tasks, to balance performance and energy efficiency.

A common hybrid approach involves using an SNN as a lightweight and efficient backbone for extracting features from event data, which are then fed into an ANN-based head for final object detection tasks. This architecture aims to achieve performance comparable to pure ANNs while significantly reducing the number of parameters, latency, and power consumption. The driving force is to bridge the gap between SNNs’ efficiency and ANNs’ superior accuracy and robust training for complex tasks. For instance, Liu et al.[102] proposed a Spiking-YOLO model for mobile robot object detection by leveraging DNN-to-SNN conversion for energy efficiency on neuromorphic hardware like TrueNorth. Similarly, Zhang et al.’s[124] spiking RetinaNet combines an SG ResNet backbone with an ANN detection head, achieving 0.296 mAP on MSCOCO. Another notable example is the DT-LIF Based SSD[68], which utilizes the DT-

LIF neuron model within a hybrid SNN based on the SSD framework. This model significantly improves detection accuracy and inference speed by employing spiking VGG and spiking DenseNet backbones, along with batch normalization (BN), spiking convolutional (SC) layers, and DT-LIF neurons, demonstrating a 25.2% improvement in object detection accuracy on the Prophesee GEN1 dataset.

The success of hybrid architectures underscores a critical insight: SNNs and ANNs should often be viewed as complementary modules, with SNNs excelling at low-level event feature extraction and ANNs at high-level classification and regression. Their synergistic combination, as exemplified by hybrid SNNs, achieves an optimal balance between performance and energy efficiency. This modular design philosophy can be readily extended to design more intricate heterogeneous systems, potentially deploying SNN components on neuromorphic chips and ANN components on GPUs for optimized overall system performance.

Attention mechanisms and transformer-based SNN Integration

The revolutionary success of transformer architectures across diverse computer vision domains has naturally led to their integration into SNNs. As presented in Fig. 6, this integration primarily aims to address the limitations of spiking CNNs in processing global context and long-range dependencies[126]. This convergence indicates that SNN development is proactively embracing and adapting to the latest advancements in modern deep learning, striving for comprehensive competitiveness.

Spiking vision transformer (S-ViT)

Spiking vision Transformers (S-ViTs) are adaptations of the vision transformer architecture for SNNs, focusing on reducing the number of timesteps for processing. The core motivation is to capture global dependencies efficiently, a known limitation of traditional SNN-CNNs, while maintaining or improving latency and energy efficiency. This is crucial for handling complex visual tasks that require a broader understanding of the image context. These models adapt the self-attention mechanism to operate with spikes, aiming to capture global dependencies efficiently. While powerful, training convergence and stability in deeper S-ViTs remain significant research challenges. Active research explores S-ViTs to improve latency and energy efficiency[127−133]. Yu et al.[134] introduced SpikingViT, a multi-scale spiking vision transformer model for event-based object detection, which enhances spatio-temporal information processing through a multi-stage feature extraction (MFE) module and a temporal memory spiking neuron (TMSN) block.

Spike-TransCNN architectures

Spike-TransCNN architectures are hybrid designs that combine spiking Transformers with spiking CNNs. These models address the inherent bias of SNN-CNNs towards local features and their difficulty in integrating global and high-level semantic information. The aim is to effectively integrate both global and multi-scale local features for enhanced detection accuracy and energy efficiency, particularly for sparse event data. They achieve this by adeptly combining the global information extraction capabilities of spiking Transformers with the local feature extraction advantages inherent in spiking CNNs. For example, in Wang et al.[103], a PSSD-transformer is proposed for image semantic segmentation, incorporating pure sparse self attention (PSSA) and dynamic spiking membrane shortcut (DSMS) to handle floating-point computations with sparse spikes. Nonetheless, potential limitations may include restricted architectural innovation and suboptimal performance on very large datasets compared to ANN counterparts.

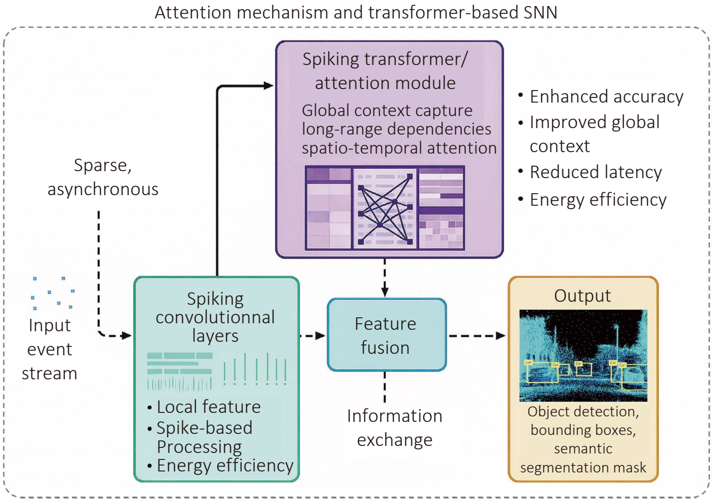

Fig. 6 | Attention mechanism and Transformer-based SNNs. This diagram illustrates how spiking versions of self-attention are integrated into SNN architectures, allowing for the efficient capture of global dependencies while preserving the energy efficiency of spike-based computation.

Other attention mechanisms

Beyond full transformer integration, various attention mechanisms have been incorporated into hybrid SNN-ANN backbones. Attention-based SNN-ANN bridging modules are designed to capture sparse spatio-temporal relationships within SNN layers and efficiently convert them into dense feature maps for the ANN component. This allows for targeted information flow and highlights salient features while maintaining SNN efficiency. The goal is to improve the interpretability and feature weighting within SNNs for better performance. Bulzomi et al.[121] proposed visual attention mechanisms for lightweight SNNs for object detection to filter noise and reduce activations. Miao et al.[77] employed fast fourier convolution (FFC) in SpikingYOLOX to provide a global receptive field, implicitly functioning as an attention mechanism for enhanced feature extraction. Zhang et al.[124] introduced an attention spike decoder (ASD) to dynamically assign weights to spiking signals along temporal, channel, and spatial dimensions for effective decoding. Fan et al.[119] integrated a multi-scale hierarchical spiking residual attention network within SFDNet. Furthermore, in Li et al.[96], specialized modules like the adaptive temporal weighting (ATW) injector, event-driven sparse (EDS) injector, and channel selection fusion (CSF) module facilitate robust interaction between SNN and ANN branches. The ATW Injector integrates event temporal features into frame features, the EDS Injector combines sparse event data with rich frame features, and the CSF module selectively fuses features from both branches. In Ma et al.[91], an Analog CBAM (convolutional block attention module[135]) combines channel and spatial attention mechanisms, designed to handle floating-point signals from ANNs to integrate attention mechanisms without corrupting spike distributions. Lastly, Wang et al.[125] introduced a dynamic surrogate gradient function (EvAF) and a boundary region-guided loss to optimize training, which involves attention to relevant boundary areas.

The integration of transformer architectures serves as a direct response to the challenges SNNs face in handling global context and long-range dependencies. A recognized weakness of SNN-CNNs is their inherent bias towards “local and single-scale features,” as well as their difficulty in “integrating global and high-level semantic information”. Transformers, by their nature, excel at capturing global attention mechanisms and long-range dependencies. While S-CNNs demonstrate robust performance in local feature extraction, they often lack sufficient global contextual information, which is paramount for many object detection tasks. The introduction of transformers into SNNs (e.g., S-ViT, Spike-TransCNN) directly compensates for the limitations of S-CNNs, thereby enhancing performance on complex visual tasks.

Multi-scale feature fusion in SNN detectors

Robust object detection, especially in real-world scenarios characterized by objects of diverse sizes, critically depends on effective multi-scale feature fusion[136], with the pipeline showed in Fig. 7. Emerging SNN architectures are increasingly incorporating sophisticated multi-scale processing capabilities. Multi-scale feature fusion is a pivotal solution to the challenge of scale variation in object detection, moving beyond the limitations of single-scale feature extraction.

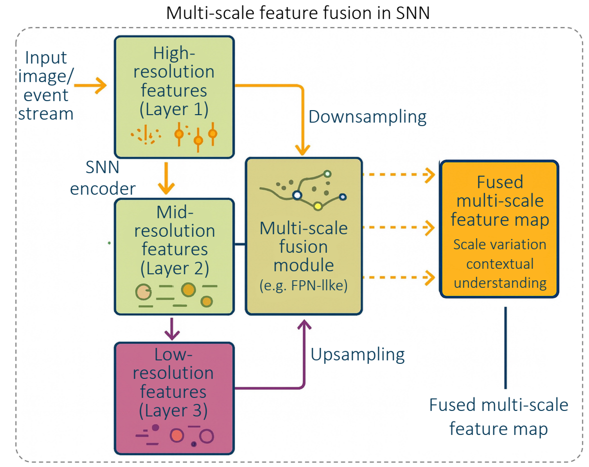
Multi-scale feature fusion in SNN
Input image/ event stream
High-resolution features (Layer 1)
Downsampling
SNN encoder
Mid-resolution features (Layer 2)
Multi-scale fusion module (e.g. FPN-llke)
Fused multi-scale feature map
Scale variation contextual understanding
Low-resolution features (Layer 3)
Upsampling
Fused multi-scale feature map

Fig. 7 | Multi-scale feature fusion in SNN. This diagram illustrates how features extracted at different resolutions are integrated within SNN architectures to robustly detect objects across a wide range of sizes.

Hierarchical feature integration

This approach involves integrating features extracted from different depths of the network. The multi-scale spiking detector framework[117] pioneered this concept to enable a hierarchical understanding of the visual scene, which is essential for detecting objects across various scales. The integration of feature pyramid structures, such as feature pyramid networks (FPNs)[136], has become a prevalent strategy in SNNs to facilitate multi-scale feature extraction[137]. This approach is directly inspired by successful methodologies in ANNs and enables SNNs to process information at multiple resolutions, enhancing the detection of objects across a wide range of scales. Zhang et al.[79] integrated an FPN architecture within their SpikeFPN for automotive event-based object detection, enhancing multi-scale feature extraction.

Spiking fusion modules

Spiking fusion object detector (SFOD)[118] and similar fusion mechanisms are being developed to integrate spike features from different scales. The goal is to ensure a comprehensive object representation, especially for dynamic objects, by combining both spatial and temporal cues from various scales. This addresses the challenge of accurately detecting moving targets where both their shape and motion patterns are crucial. SFOD's spiking fusion module not only achieves spatial fusion but also enhances multi-scale features in the temporal domain, allowing the integration of temporal cues (e.g., motion patterns) from different scales to improve detection accuracy for dynamic objects. Fan et al.[120] integrated the spiking multi-direction fusion module (SMFM) within SpikeDet to enhance multi-scale feature fusion and preserve neuron firing patterns for object detection. Miao et al.[77] integrated SPP-SNN (spatial pyramid pooling spiking neural network) within their SpikingYOLOX to enhance multi-scale feature fusion capabilities for object detection. Qu et al.[97] addressed limitations in SPPF (spatial pyramid pooling fast) conversion by introducing a spike-maxpooling mechanism, enabling lossless conversion and enhancing multi-scale feature fusion in ultralow-latency SNNs. Fan et al.[119] developed a lightweight spiking aggregation module within SFDNet for efficient RGB-event fusion in object detection.

Lightweight multi-fusion architectures

SLP-Net (lightweight multi-fusion UNet based on spiking neural systems)[138] introduces a multi-channel SNP-type convolution (MCConvSNP) neuron model within a lightweight asymmetric encoder-decoder design. This architecture aims for high accuracy with low parameters and FLOPs, particularly for tasks like skin lesion segmentation, by optimizing feature extraction and fusion across multiple levels. It uses an efficient multi-scale feature extraction block (EMFE) with dilated convolutions for deep stage multi-scale feature extraction, and a multi-level feature fusion module (MFF) in skip connections for hierarchical fusion. A spatial-channel fusion module (SCF) further optimizes feature fusion across spatial and channel dimensions. Wang et al.[125] utilized a spiking multi-scale context aggregation (SMSCA) module to fuse features from different scales and enhance contextual information. The module obtains multiple scales of receptive fields through average pooling with different kernel sizes and strides. Ye et al.[80] introduced spiking-SegNet for image segmentation, employing a U-shaped full-convolutional architecture with a spiking encoder-decoder that extracts multi-scale information via convolutional and transposed convolutional layers.

The inherent ability of SNNs to process spatio-temporal information means that their multi-scale fusion extends beyond mere spatial dimensions; it can also integrate information across different layers and time points, forming a "temporal perception" across scales. This spatio-temporal multi-scale fusion capability represents a unique advantage for SNNs in dynamic object detection, potentially leading to more precise and real-time detection of moving objects than traditional ANNs.

SNN architectures for image semantic segmentation

Semantic segmentation, a crucial computer vision task, aims to assign a class label to every pixel in an image, facilitating a fine-grained understanding of the scene[47]. Unlike object detection, semantic segmentation typically does not distinguish between individual instances of the same object category, instead treating all pixels belonging to a class (e.g., all pixels of "road") as a single entity[55]. Traditional approaches to semantic segmentation predominantly rely on deep learning methods, particularly CNN architectures based on encoder-decoder structures like FCN[47], SegNet[139], U-Net[140], DeepLab[141], and PSPNet[142]. Among these, U-Net, with its iconic U-shaped structure and prominent skip connections, has achieved significant success particularly in medical image segmentation[143].

Semantic segmentation finds extensive applications in critical domains such as autonomous driving, drone navigation, medical image analysis, and augmented reality. While traditional CNNs have demonstrated commendable performance in semantic segmentation, they face inherent tradeoffs concerning latency, accuracy, and energy efficiency, particularly in real-time systems like autonomous vehicles and drones[55]. SNNs, with their low-power and event-driven characteristics, offer a novel avenue to circumvent these bottlenecks. The substantial computational demands of traditional CNNs often render them unsuitable for deployment on edge devices requiring real-time performance and high efficiency. Consequently, SNNs' low-power and low-latency attributes position them as a promising solution for deploying semantic segmentation in resource-constrained environments.

Event cameras, which capture dynamic scene changes by outputting asynchronous streams of brightness changes, provide data with high temporal resolution, low latency, and low power consumption. This data modality aligns exceptionally well with the requirements of semantic segmentation for dynamic scene understanding and real-time processing, offering a natural advantage for SNN applications in this domain. Semantic segmentation demands the capture of dynamic information; event cameras provide high-temporal-resolution event streams that SNNs, as event-driven networks, can process with superior efficiency and potential compared to frame-based ANNs, especially for dynamic semantic segmentation tasks.

Unique advantages of SNNs in semantic segmentation

The inherent event-driven nature, low power consumption, minimal latency, and native spatio-temporal information processing capabilities of SNNs render them exceptionally well-suited for handling data originating from event cameras. This positions SNNs to exhibit substantial potential in real-time semantic segmentation[86]. When deployed on specialized neuromorphic hardware, SNNs can achieve ultra-low power consumption and latency, which is indispensable for cutting-edge edge AI applications. SNNs communicate through binary spike signals, enabling them

Table 2 | Key SNN architectures for object detection.

Architecture type	Specific model name	Quantitative metrics	Energy efficiency	Relevant datasets	Limitations / Challenges
Pure SNN (S-CNN)	Spiking-YOLO[76]	N/A	N/A	Static & dynamic datasets	High training difficulty, persistent accuracy limitations
	Spiking-YOLOX[77]	State-of-the-art mAP	Low computational requirements	N/A	Complex optimization, potential performance degradation
	YOLO-C3 SNN[116]	Improved accuracy (no specific mAP)	2.47x lower power consumption	Overhead power lines	Needs direct training with surrogate gradient
	Embedded SNN[84]	Small mAP loss	490 mJ/prediction	Automotive event data	Requires post-training quantization (PTQ)
	Spiking CenterNet[78]	2.6% higher mAP	Better power efficiency	N/A	Requires KD from non-spiking teacher
	SpikiLi: LiDAR 3D SNN[95]	CNN-comparable precision	Low inference latency (3ms)	LiDAR-based 3D	Primarily simulation-based, needs hardware validation
	Multi-patch SNN[122]	98.9% accuracy	0.336W power (20FPS)	Infrared drone detection	Relies on ANN-SNN conversion, may have latency
	Sparse compressed SNN Accelerator[101]	N/A	1.05 mJ/frame	N/A	Primarily hardware-focused, limited generalizability
	EMS-YOLO[123]	ANN-comparable performance	5.83x less energy (4 timesteps)	N/A	May still have accuracy gap for very large datasets
	BN-SNN with BIF neurons[87]	0.476 mAP@0.5 (MS-COCO), 0.591 (GEN1)	Reduced temporal steps	MS-COCO, GEN1	Primarily conversion-based, may have some conversion loss
	DT-LIF based SSD[68]	25.2% accuracy improvement	N/A	Prophesee GEN1	General SNN training difficulties
Attention mechanism or transformer-based SNN	Direct training high-performance SNNs[124]	0.296 mAP	N/A	MSCOCO	Requires direct training, may not match all ANN performance
	Tiny SNN with visual attention[121]	N/A	Energy-efficient on SpiNNaker	N/A	May require dynamic weight adjustment
Multi-scale feature fusion	MSD (multi-scale spiking detector)[117]	High performance	Low power	N/A	Further exploration of effective model construction needed
	SFOD (spiking fusion object detector)[118]	Improved detection accuracy	N/A	Dynamic objects	Complexity in module design, reliance on specific data modalities
	Spike-BRGNet[125]	+1.57-1.91% mIoU	17.76x less energy than ANN	DDD17, DSEC	Relies on event data (no image frames)
	SpikeFPN[79]	0.477 mAP	Energy-efficient	GEN1 (automotive event data)	Relies on direct training with SG
	SUHD: Ultralow-latency SNN[97]	30% mAP enhancement	Ultralow latency	N/A	Complex conversion, may rely on specific datasets

to replace the high-power multiply-accumulate (MAC) operations prevalent in traditional ANNs with more energy-efficient accumulate (AC) operations, thereby significantly enhancing energy efficiency.

The benefits of SNNs in semantic segmentation extend beyond mere energy efficiency (by avoiding unnecessary computations) to encompass the low latency afforded by their event-driven and asynchronous computation. This is critically important for real-time applications like autonomous driving, which necessitate rapid decision-making[55]. Semantic segmentation in domains such as autonomous driving demands real-time responsiveness. SNNs' event-driven and sparse computational paradigm provides low power consumption and low latency, enabling them to meet the stringent real-time requirements of semantic segmentation. The synergy between SNNs and event cameras allows them to demonstrate unparalleled — and potentially irreplaceable — competitiveness in semantic segmentation tasks under extreme conditions (e.g., high-speed motion, high dynamic range, low light) where traditional frame-based cameras struggle. While traditional frame cameras are limited in these challenging environments, event cameras excel at capturing dynamic changes. SNNs' natural compatibility with event cameras allows them to provide more refined and real-time scene understanding than traditional ANNs in these specific scenarios, potentially even surpassing them.

Encoder-decoder and U-Net-like SNN architectures

Semantic segmentation tasks commonly employ encoder-decoder architectures, where an encoder extracts high-level semantic features and downsamples spatial resolution, and a decoder upsamples to restore the original resolution, generating dense pixel-wise segmentation masks[47]. U-Net, with its iconic U-shaped structure and prominent skip connections, stands as a prime example of such architectures, achieving significant success particularly in medical image segmentation[140]. SNNs have strategically adopted this highly successful encoder-decoder paradigm, adapting it to their unique characteristics to maximize efficiency and performance. SNNs that adopt U-Net-like encoder-decoder structures for semantic segmentation are not merely direct copies but rather undergo strategic "lightweight modifications" and "spiking transformations" tailored to SNN characteristics.

Event-data optimized SNN-UNet

EvSegSNN[86] is a bio-inspired encoder-decoder SNN architecture explicitly designed with a U-Net layout and optimized for event data. It integrates parameterized leaky-integrate-and-fire (PLIF) neurons and lightweight modifications to the U-Net structure, reducing depth and convolutional layers. The primary motivation is to effectively process high-temporal-resolution, high-dynamic-range, and low-latency asynchronous spike events from event cameras, reducing parameter count while maintaining performance for real-time applications. These methods significantly cut down parameters (8.55 million vs. 13.46 million baseline). On the DDD17 dataset, EvSegSNN achieved 45.54% MIoU and 89.90% accuracy[86].

Efficient spiking encoder-decoder networks

Spiking encoder-decoder network (SpikingEDN)[64] is an efficient spiking encoder-decoder network specifically developed for large-scale event-based semantic segmentation tasks. The aim is to optimize spiking operations for dense prediction, addressing the challenge of efficiently processing large-scale event data while achieving competitive semantic segmentation performance. Its design achieves an impressive 72.57% MIoU on the DDD17 dataset and 58.32% on the DSEC-Semantic dataset, demonstrating competitive performance against state-of-the-art ANNs with significantly reduced computational demands.

Lightweight transformer-based SNN for segmentation

Spike-driven lightweight transformer-based semantic segmentation network (SLTNet)[144] introduces a hierarchical single-branch SNN with an encoder-decoder framework. The encoder uses spike-driven convolutional blocks (SCBs) for local detail and spike-driven transformer blocks (STBs) for long-range context. SLTNet aims to fully capitalize on SNNs' strengths, particularly the low computational cost of its SCBs and STBs, while leveraging transformer-like mechanisms to capture global context, a common challenge for SNN-CNNs. A lightweight spiking decoder recovers spatial details via feature fusion, and its spike-LD module enables multi-scale feature capture. On the DDD17 and DSEC-Semantic datasets, SLTNet achieved a significant mIoU improvement, reduced energy consumption by 4.58 times, and boasted an inference speed of 114 FPS, with substantially lower parameter counts and FLOPs compared to existing methods.

Spiking-UNet architectures

Chakravarty et al.[81] proposed a modified U-net architecture within an SNN framework tailored to operate without dense layers, producing segmented images as output. Dakic et al.[82] utilized a Spiking-UNet architecture with LIF neurons and constant current injection encoding for spectrum occupancy monitoring. Li et al.[89] introduced a Spiking-UNet for image processing, combining SNNs with the U-Net architecture. These models adapt the highly successful U-Net paradigm to SNNs to leverage its effective encoder-decoder structure for segmentation tasks while aiming for SNN's characteristic energy efficiency and low latency. Chakravarty et al.'s model uses approximated gradients. Dakic et al.'s model employs constant current injection encoding and BCE+Dice loss, achieving a 10x energy reduction compared to CNNs. Li et al.'s model addresses information propagation and training challenges by proposing multi-threshold spiking neurons and a conversion/fine-tuning pipeline with connection-wise normalization, reducing inference time by approximately 90%.

Transformer and NSNP-based SNNs for segmentation

Lei et al.[88] introduced Spike2Former, an efficient spiking Transformer for image segmentation, adapted from Mask2Former, using spike-driven deformable transformer encoder (SDTE) and spike-driven mask embedding (SDME). Sun et al.[145] proposed NSNPFormer, a Transformer-based semantic segmentation method integrating the convolutional nonlinear spiking neural P (NSNP) model. These architectures seek to overcome the limitations of traditional SNN-CNNs in capturing global context and long-range dependencies, essential for accurate and complex image segmentation. They also aim to enhance information representation and training stability in deep SNNs. Spike2Former uses normalized integer LIF (NI-LIF) neurons for training stability and achieved state-of-the-art accuracy across ADE20K, VOC2012, and CityScapes datasets with significant energy efficiency. NSNPFormer uses parallel ConvNSNP and transformer channels with residual connections for local and global feature extraction, achieving mIoU scores of 53.7% on ADE20K and 58.06% on Pascal Context datasets.

Specialized SNNs for medical image segmentation

Li et al.[146] developed ODCS-NSNP, a deep segmentation network for optic disc and cup segmentation based on nonlinear spiking neural P systems. Yang et al.[138] proposed SLP-Net for skin lesion segmentation, introducing a multi-channel SNP-type convolution (MCConvSNP) neuron model and a lightweight asymmetric encoder-decoder design. These models address the critical need for efficient and accurate segmentation in medical imaging, where precision and computational efficiency are paramount for diagnostic tools. ODCS-NSNP features a densely connected depth-separable network unit (SDN-Unit) and a redesigned resampling operator (SRS-Operator) to improve boundary accuracy and multi-scale feature extraction. SLP-Net utilizes EMFE, MFF, and SCF modules for feature fusion, achieving high accuracy with low parameters and FLOPs.

Advanced spiking-SSegNet and analog SNN-UNet

Ye et al.[80] proposed Spiking-SSegNet, a U-shaped full-convolutional semantic segmentation network built on the Spiking-NSNet model, utilizing a hybrid attenuation factor setting. Ma et al.[91] proposed Analog Spiking U-Net (AS U-Net), which integrates Analog CBAM and Spiking ViT modules into an SNN framework. These models aim to improve mIoU with low latency and minimize information loss, respectively, pushing the boundaries of SNN performance in semantic segmentation tasks. Spiking-SSegNet leverages the hybrid attenuation factor for improved mIoU with low latency. AS U-Net adjusts neuron firing positions to transfer information as floating-point signals, minimizing information loss, and achieved high accuracy on diabetic retinal vessel segmentation datasets, demonstrating SOTA energy efficiency.

Although SNNs historically faced performance challenges in dense prediction tasks like semantic segmentation, the advent of models such as SpikingEDN and SLTNet, through optimized architectural design and effective utilization of event data properties, have achieved “competitive” or even “superior” performance compared to state-of-the-art ANNs[144]. This marks a significant maturation of SNNs in complex visual tasks, indicating broad and promising prospects for their application in dense prediction fields.

Hybrid SNN-ANN/transformer architectures in semantic segmentation

To fully capitalize on the respective strengths of ANNs and SNNs, researchers have extensively explored hybrid architectures for semantic segmentation. In these configurations, SNNs are often employed in the encoder section to process event data efficiently, while ANNs are utilized in the decodersectionforrobustreconstructiontasks.Thisburgeoning interest in hybrid architectures reflects a broader trend within the AI landscape, shifting from the pursuit of a singular, optimal model towards the exploration of multi-paradigm, heterogeneous computing solutions.

Event-frame fusion hybrid framework

The event-frame fusion hybrid framework[96] combines an SNN branch for event data and an ANN branch for frame data to leverage complementary information from both modalities. Existing event-based semantic segmentation methods frequently fail to leverage the complementary information provided by both event and frame data. A single event stream may lack crucial visual detail, while traditional frame data processing is computationally expensive. This framework addresses these limitations by integrating both modalities for comprehensive scene understanding. Specialized modules like the adaptive temporal weighting (ATW) injector (dynamically integrating event temporal features into frame features), event-driven sparse (EDS) injector (combining sparse event data with rich frame features), and channel selection fusion (CSF) module (selective feature fusion) facilitate robust interaction and information exchange between branches, aiming for comprehensive and accurate scene understanding.

Hybrid spiking fully convolutional neural networks

The hybrid SFCNN (spiking fully convolutional neural network)[147] employs a hybrid architecture for semantic segmentation, leveraging binary information transmission in its encoder. This architecture seeks to combine the energy efficiency of SNNs with the robust learning capabilities of FCNs for dense prediction tasks, overcoming the limitations of pure SNNs in complex segmentation. It uses a surrogate gradient method for direct backpropagation training. On the VOC2012 dataset, this model achieved a significant mIoU improvement (almost 30% higher than existing spiking FCNs), demonstrating the feasibility of end-to-end optimization.

NSNPFormer with transformer integration

NSNPFormer[145] integrates the convolutional nonlinear spiking neural P (NSNP) model with transformers for semantic segmentation. This model addresses the need for both local feature extraction (using ConvNSNP) and global contextual information capture (using transformers), which are crucial for accurate semantic segmentation. It features parallel ConvNSNP and transformer channels with residual connections, enabling efficient local feature extraction and global contextual information. NSNPFormer achieved notable mIoU scores on ADE20K and pascal context datasets.

Spiking-LSTM for hyperspectral image segmentation

The Spiking-LSTM model[104] combines SNN and LSTM architectures for hyperspectral image segmentation. This architecture is designed to effectively capture spatio-temporal information inherent in hyperspectral images, enabling accurate early-stage detection of plant diseases like Sclerotinia, while maintaining energy efficiency. It replaces traditional LSTM gating functions with spiking neurons and uses surrogate gradients for backpropagation. The model achieves 94.3% mAP for Sclerotinia detection on rapeseed leaves, extracting early infected areas. It demonstrates high accuracy with significantly lower energy consumption (one-fifth of traditional LSTM), highlighting its potential for efficient disease detection.

This burgeoning interest in hybrid architectures reflects a broader trend within the AI landscape, shifting from the pursuit of a singular, optimal model towards the exploration of multi-paradigm, heterogeneous computing solutions. Future research will likely witness the emergence of even more intricate hybrid models, not only combining different neural network types but also integrating across distinct hardware platforms. For instance, SNN components might be deployed on neuromorphic chips, while ANN components reside on GPUs, optimizing overall system performance.

Multi-scale feature fusion and contextual dependency handling

Multi-scale representations are fundamental for accurately segmenting objects of varying sizes within an image. Furthermore, effectively capturing long-range pixel dependencies and channel-wise feature similarities is crucial for enhancing pixel-level region understanding and improving the overall performance of segmentation models. SNNs address these critical challenges in semantic segmentation through several innovative approaches.

SLTNet's multi-scale modules

SLTNet's Spike-LD module[144] introduces a novel three-branch structure that integrates dilated and depthwise separable convolutions. Its spike-driven transformer blocks (STBs) are specifically designed to bolster long-range contextual feature interactions. These modules are essential for capturing multi-scale features across different receptive fields and for efficiently processing and fusing information at various scales, which is critical for enhancing segmentation accuracy. STBs address the need for global spatial relationships, a weakness of traditional CNNs. The spike-LD module enables simultaneous processing and fusion of information at various scales. STBs use a spike-driven multi-head self-attention module (SDMSA) to efficiently capture global spatial relationships, augmented by a multilayer perceptron (MLP) for channel-wise information. SDMSA effectively reduces computational complexity to $O(N)$ (where N is the number of tokens/features), and the entire block primarily involves floating-point addition operations, leading to significant reductions in energy consumption and improved efficiency .

Feature enhancement and aggregation

Within the decoder of certain SNN segmentation architectures, a feature enhancement (FE) module restores fine spatial details and integrates features from different hierarchical levels. EMSNet (enhanced multi-scale networks)[148] employs an integration of enhanced regional module (IERM) and multi-scale convolution module (MSCM). These modules are crucial for producing refined segmentation masks by integrating fine-grained details from early layers with high-level semantic information, and for robustly handling objects of varying scales. The FE module contributes to a more refined processing of multi-scale information. IERM enhances fused feature representation through dynamic convolutional structures, while MSCM gathers multi-scale contextual information using deformable deep convolutions and multi-branch deep asymmetric convolutions.

Spiking neural P systems for multi-scale features

ODCS-NSNP[146] introduces a redesigned resampling operator (SRS-Operator) based on ConvSNP that resamples multiple features from large regions into multiple output features. Spike-BRGNet[125] includes a spiking multi-scale context aggregation (SMSCA) module. These systems aim to capture long-term dependencies and preserve fine spatial details, improving segmentation boundary accuracy, especially for complex anatomical structures in medical images, or for robust scene understanding in traffic environments. The SRS-Operator, based on ConvSNP, captures long-term dependencies and preserves fine spatial details. The SMSCA module aggregates features from five different scales using average pooling and BN-LIF-Conv processing, enabling the network to capture global contextual information and multi-scale features for accurate segmentation.

Multi-scale feature fusion is a pivotal solution to the challenge of scale variation in object detection. Traditional CNNs, often limited to local and single-scale features, inherently suffer from restricted detection accuracy. SNNs are actively overcoming these limitations by adopting successful strategies from ANNs, such as dilated convolutions and Transformer-like mechanisms, and adapting them for spike-based computation[144]. SLTNet's ablation studies confirm the complementary roles of transformer-like and convolutional blocks in SNNs, which collectively enhance performance[144]. Given SNNs' inherent capacity to process spatio-temporal information, their multi-scale feature fusion and contextual dependency handling extend beyond mere spatial dimensions to integrate information across the temporal axis. This ability is particularly advantageous for understanding semantic changes in dynamic scenes (e.g., the movement of obstacles in autonomous driving), potentially enabling SNNs to provide more precise and real-time detection of moving objects than traditional ANNs.

Learning algorithms and training strategies for SNNs

The training of SNNs presents one of the most significant challenges for their widespread adoption, primarily due to the inherent discrete and non-differentiable nature of spike generation, which renders traditional gradient-based optimization methods, such as backpropagation, directly inapplicable. To overcome these fundamental difficulties, a diverse array of innovative learning algorithms and training strategies are developed.

Surrogate gradient methods

The core challenge in training SNNs with gradient-based methods stems from the spike generation function, which is typically a Heaviside step function. This function can be

Table 3 | Key SNN architectures for semantic segmentation.

Architecture type	Specific model name	Quantitative metrics	Energy efficiency	Relevant datasets	Limitations / Challenges
Pure SNN	EvSegSNN[86]	45.54% MIoU, 89.90% Accuracy	Reduced parameters	DDD17	Higher average firing rate than ideal sparse SNNs
	SpikingEDN[64]	72.57% MIoU (DDD17), 58.32% MIoU (DSEC-Semantic)	Reduced computational resources	DDD17, DSEC-Semantic	Training stability, generalization on complex scenes
	Spiking-UNet (Seg)[81]	99% DSC	N/A	EM segmentation 2015, Data Science Bowl 2018	Encoding real-life datasets is difficult
	Spiking-UNet (Spectrum)[82]	Similar TPr to CNN	10x energy reduction vs CNN	Spectrum monitoring	Optimization process intricate
	Spiking-UNet (Multi-threshold)[89]	Comparable to non-spiking U-Net	90% inference time reduction	N/A	Complex to determine optimal thresholds, may overfit
	Early SNN (Unsupervised)[114]	Basic pattern classification	N/A	N/A	Limited to simple tasks, fixed parameters
	Spiking-SSegNet[80]	43.2% mIoU (PASCAL VOC2012), 53.4% mIoU (DDD17)	Low latency (2 time steps)	PASCAL VOC2012, DDD17	No explicit limitations mentioned
Hybrid SNN	Event-frame fusion Hybrid framework[149]	Improved accuracy	Improved efficiency	N/A	Complex training strategies, increased computational cost
	Hybrid SFCNN[147]	30% mIoU improvement over SNN FCNs	N/A	VOC2012	Training difficulty persists, performance needs improvement
	NSNPFormer[145]	53.7% mIoU (ADE20K), 58.06% mIoU (Pascal Context)	N/A	ADE20K, pascal context	Relies on ResNet backbone, limited to specific datasets
	Spiking-LSTM[104]	94.3% mAP	1/5th of conventional LSTM energy	Rapeseed leaves (Sclerotinia)	Requires multiple simulation steps, may not exceed 90% mAP
Attention mechanism or transformer-based SNNs	Spike2Former[88]	SOTA accuracy	5.0x-6.6x efficiency	ADE20K, VOC2012, CityScapes	Still underperforms ANNs in complex tasks, challenges in deep S-ViTs
	SLTNet[144]	Significant mIoU improvement	4.58x energy reduction, 114 FPS	DDD17, DSEC-semantic	Lacks visual detail, computation can be expensive or rely on auxiliary images
	Analog spiking U-Net (AS U-Net)[91]	90.4% mIoU, 98.3% PixAcc	SOTA energy efficiency	Diabetic retinal vessel segmentation	No explicit limitations mentioned
Multi-scale feature fusion	SLP-Net[138]	93.87% Acc, 88.21% DSC	Low parameters (0.1M), fast processing	ISIC2018 (skin lesion)	Asymmetric design may cause information loss, relies on supervised data
	ODCS-NSNP[146]	0.9817 Dice (OD), 0.9859 Dice (OC) (RIM-ONE-r3)	N/A	RIM-ONE-r3, Drishti-GS, REFUGE	No explicit limitations mentioned

defined as

$$S\left(V_{\mathrm{m}}\right)=H\left(V_{\mathrm{m}}-V_{\mathrm{th}}\right)=\begin{cases}1 & \text{if } V_{\mathrm{m}} \geq V_{\mathrm{th}} \\ 0 & \text{if } V_{\mathrm{m}} < V_{\mathrm{th}}\end{cases} . \quad (5)$$

The derivative of the heaviside function, $\mathrm{d}S/\mathrm{d}V_{\mathrm{m}}$, is zero everywhere except at the threshold $V_{\mathrm{m}}=V_{\mathrm{th}}$, where it is undefined (an impulse or dirac delta function). This property effectively blocks gradient propagation during backpropagation. To circumvent this issue, surrogate gradient (SG) methods have been proposed[150−151]. The central idea of SG is to approximate the non-differentiable derivative of the heaviside function with a continuous and differentiable function during the backward pass (gradient calculation), while keeping the forward pass (spike generation) intact.

Specifically, for the backward pass, instead of computing $\mathrm{d}S/\mathrm{d}V_{\mathrm{m}}$, a smooth, differentiable surrogate derivative function $\sigma'\left(V_{\mathrm{m}}\right)$ is used:

$$\frac{\partial L}{\partial V_{\mathrm{m}}}=\frac{\partial L}{\partial S}\cdot\sigma'\left(V_{\mathrm{m}}\right) , \quad (6)$$

where L is the loss function. Common choices for the surrogate derivative function $\sigma'\left(V_{\mathrm{m}}\right)$ include[152]:

1) Rectangular (e.g., identity function for a limited range): simplest approximation, where $\sigma'\left(V_{\mathrm{m}}\right)$ is a constant or linear function within a narrow window around the threshold and zero elsewhere.

2) Sigmoid derivative: the derivative of the sigmoid function, $\sigma'(x) = \sigma(x)(1-\sigma(x))$, provides a bell-shaped curve.

3) Arctan derivative: a widely used and effective surrogate gradient function is the derivative of the arctangent function. For a membrane potential V_{m} and threshold V_{th}, a common form is

$$\sigma'(V_{\mathrm{m}}) = \frac{1}{\pi}\frac{1}{1+\left(\dfrac{V_{\mathrm{m}}-\mu}{\alpha}\right)^2}, \tag{7}$$

where μ is typically set to V_{th}, and α is a scaling factor that controls the width of the peak. This function exhibits a smooth, bell-shaped peak around the threshold, allowing gradients to propagate effectively when the neuron's membrane potential is close to firing. As the potential moves away from the threshold, the gradient smoothly decays to zero, preventing issues like vanishing or exploding gradients.

4) Piecewise exponential (PiecewiseExp) and Gaussian error: Zhang et al.[104] compared PiecewiseExp and Erf as surrogate gradient functions for training Spiking-LSTM, finding PiecewiseExp to consistently yield better detection accuracy and stability.

5) Evolutionary asymptotic function (EvAF): Wang et al.[125] introduced EvAF as a dynamic surrogate gradient function that replaces infinite gradient values with real numbers, allowing effective weight updates during initial training and improving accuracy in backward gradient computation.

Recent works extensively employing surrogate gradient methods include direct training of deep SNNs for object detection[123–124], SpikeFPN[79], SFDNet[119], LT-SNN with Separate Gradient Path[85], SUHD[97], and Hybrid SFCNN for semantic segmentation[147]. Chakravarty et al.[81] explored modern backpropagation techniques in SNNs, focusing on surrogate or approximate gradient methods to overcome non-differentiable functions. Lei et al.[88] employed surrogate gradients to approximate derivatives of non-differentiable spike functions, enabling training of complex SNN architectures for image segmentation. Ye et al.[80] used spatio-temporal backpropagation (STBP) for direct SNN training, fusing spatial and temporal domains and addressing non-differentiability with triangular surrogate gradients.

SG methods enable SNNs to be trained using established deep learning frameworks like Backpropagation Through Time (BPTT)[153–155] or Spatio-Temporal BackproPagation (STBP)[156–157], thus overcoming a core impediment to deep SNN adoption. While practical, SGs are approximations, sometimes leading to accuracy degradation or convergence issues. Future research aims to develop more precise and efficient SG functions or gradient-free methods to balance performance and biological realism.

Spiking batch normalization

Batch normalization (BN)[158] has been instrumental in stabilizing and accelerating deep ANN training, improving convergence and generalization. However, its direct application to SNNs poses significant challenges due to the sparse and discrete nature of SNN activations (spikes) and the complex dynamics of membrane potentials. Traditional BN, which normalizes activations to have a mean of 0 and variance of 1, can cause membrane potentials to become excessively high or low, disrupting spike firing patterns and leading to "dead neurons" (neurons that never fire) or "bursting neurons" (neurons that fire too frequently)[159].

To mitigate these issues, threshold-dependent batch normalization (tdBN) was specifically designed for SNNs[159]. The core principle of tdBN is to normalize the membrane potential of neurons not to a fixed mean and variance, but relative to their firing threshold. This ensures that the membrane potentials are kept within an optimal range, allowing for stable and balanced spike activity. The tdBN operation for a membrane potential V_{m} can be formulated as:

$$\widehat{V}_{\mathrm{m}} = \gamma\frac{V_{\mathrm{m}}-\mu_{\mathrm{B}}}{\sigma_{\mathrm{B}}}+\beta, \tag{8}$$

where μ_{B} and σ_{B} are the batch mean and standard deviation of V_{m}, and γ and β are learnable scaling and shifting parameters, similar to conventional BN. However, in tdBN, these parameters or the normalization targets are dynamically adjusted based on the neuron's threshold or desired firing rate. For instance, tdBN can normalize the membrane potential such that it fluctuates around the neuron's specific firing threshold, thereby stabilizing spike generation. This customized normalization helps maintain appropriate spiking activity levels across the network, preventing issues like vanishing or exploding gradients and significantly improving the training stability and performance of large-scale deep SNNs.

Recent works that employ tdBN include spiking-YOLOX[77], which integrates learnable decay parameters along with tdBN for computational efficiency. Zhang et al.[79]'s SpikeFPN utilizes an adaptive threshold mechanism for stable training. Yu et al.[134] employ tdBN in their SpikingViT model to ensure stable pulse signal propagation. Fan et al.[119] introduced separated batch normalization (SeBN) in SFDNet, which normalizes feature maps independently across multiple time steps and optimizes integration with residual structures to capture temporal dynamics more effectively. Li et al.[89] proposed a connection-wise normalization method for Spiking-UNet to prevent inconsistent firing rates in skip connections, ensuring accurate information representation by normalizing weights based on scale factors. Lei et al.[88] addressed training stability in complex SNN architectures by proposing Normalized Integer LIF (NI-LIF) neurons, which normalize integer activations during training to ensure precise feature representation and mitigate quantization error.

The development of tdBN exemplifies a crucial trend: instead of merely porting successful ANN techniques to SNNs, researchers are carefully adapting them to align with

SNNs' unique computational mechanisms. This tailored approach effectively addresses SNN-specific training hurdles. tdBN, alongside surrogate gradient methods, forms a critical dual pillar supporting the resolution of SNNs' training complexity, underscoring the necessity of multi-faceted, systemic solutions rather than relying on a singular algorithmic breakthrough.

Other training optimization techniques

Beyond surrogate gradients and tdBN, research is actively exploring a variety of other training optimization techniques to further enhance SNN performance and stability:

ANN-to-SNN conversion

ANN-to-SNN conversion is a widely adopted training strategy that involves first training a high-performing traditional ANN with standard backpropagation, and then converting its weights and activations into an SNN[160]. This method leverages the maturity and robust training capabilities of ANNs, though it can incur some accuracy loss and increased SNN inference time for complex tasks. Key techniques include weight normalization, which scales weights to maximize firing rates in the SNN after conversion, and threshold calibration, which adjusts neuron thresholds to match the dynamic range of ANN activations. Examples include early conversion for energy-efficient object recognition[115], Spiking-YOLO for mobile robots[102], LiDAR-based 3D object detection[95], sparse compressed SNN accelerators[101], spike calibration[161] and bistable integrate-and-fire neurons[87]. Spiking-UNet[89] adopts a conversion and fine-tuning pipeline, leveraging pre-trained U-Net models to reduce time steps while preserving performance. Lei et al.[88] address challenges in complex SNN architectures by normalizing integer spiking neurons during conversion. Ye et al.[80] introduce a transfer learning approach for Spiking-SSegNet, where pre-trained Spiking-NSNet weights are fine-tuned for semantic segmentation, improving performance and reducing training costs.

Event-driven backpropagation

This approach aims for higher biological fidelity by directly propagating errors through discrete spike events, rather than relying on surrogate gradients[162–165]. This typically involves more complex theoretical frameworks, such as event-based error propagation rules or credit assignment mechanisms sensitive to spike timings. While conceptually appealing for its biological realism, its implementation and training remain more intricate than SG methods.

Hardware-aware training

As neuromorphic hardware matures, training strategies increasingly incorporate hardware-specific constraints and advantages. This involves designing algorithms optimized for the unique parallel processing, memory architectures, and communication mechanisms of neuromorphic chips (e.g., Intel Loihi[69], IBM TrueNorth[166]), thereby maximizing SNNs' potential on these specialized platforms.

Recurrent SNNs and backpropagation through time (BPTT)

For tasks requiring sequence processing or memory, recurrent SNNs (RSNNs) are utilized. Training RSNNs often involves BPTT, where gradients are unrolled over time. While powerful, this can be computationally expensive and suffer from vanishing/exploding gradients, similar to recurrent ANNs, requiring careful regularization and optimization[153]. Chakravarty et al.[81] highlight spike-timing-dependent plasticity (STDP) and various backpropagation forms adapted for non-differentiable spike functions. Ye et al.[80] introduce a temporal correlated loss (TC) algorithm to optimize SNN direct training, ensuring faster convergence and improved robustness by adjusting neuronal membrane potential distribution at each time step.

Loss functions for SNNs

Various loss functions are employed to optimize SNN training for specific tasks. Chakravarty et al.[81] implement a modified U-net architecture that uses surrogate/approximate gradient methods to calculate gradients for the error function. Dakic et al.[82] employ a combined Binary Cross Entropy (BCE) and Dice loss function for image segmentation tasks. Li et al.[96] utilize the LCE (total BCE loss) for their event-frame fusion framework. Lei et al.[88] use categorical cross-entropy loss to fine-tune converted SNN models. Sun et al.[145] employ a cross-entropy loss function to optimize ODCS-NSNP. Yang et al.[138] utilize accuracy, sensitivity, specificity, jaccard, and dice similarity coefficient as evaluation metrics, with the final loss being a weighted sum of losses from different stages. Wang et al.[125] introduce a novel boundary region-guided loss function, combined with regular and early-stage cross-entropy semantic losses, to optimize the network. Zhang et al.[104] use a weighted cross-entropy loss function to address data imbalance in Sclerotinia detection.

The intricate nature of SNN training necessitates a multi-pronged research effort. The aforementioned methods, ranging from biologically inspired rules to clever adaptations of existing deep learning techniques, collectively form a strategic response to SNNs' training challenges. This flexibility and innovation are crucial for overcoming technical hurdles. Importantly, these methods are not mutually exclusive but often complementary, jointly accelerating the advancement of SNN training technology. While ANN-to-SNN conversion offers a pragmatic path, its inherent accuracy loss and increased inference timesteps limit its long-term viability for achieving cutting-edge performance. Consequently, the prevailing research trend is gravitating towards more efficient, end-to-end direct training methods that can fully leverage the intrinsic advantages of SNNs.

Current capabilities, challenges, and future outlook

SNNs have demonstrated immense potential in image object detection and semantic segmentation, particularly in terms of energy efficiency and real-time processing. This positions them as a pivotal technology for edge computing and autonomous systems operating under strict resource constraints.

Current performance benchmarks and application potential

In terms of architectural innovation, SNNs have made significant strides. The evolution from rudimentary S-CNNs to sophisticated hybrid SNN-ANN models, and further to the integration of transformer-based and attention mechanisms, reflects a continuous exploration by researchers for more efficient and accurate architectures. Hybrid architectures, such as the general Hybrid-SNN designs and the DT-LIF Based SSD, effectively bridge the performance gap of pure SNNs by synergistically combining SNNs' energy efficiency with ANNs' powerful representation capabilities.

Despite SNNs having historically lagged behind DNNs on certain complex tasks, the performance disparity is rapidly diminishing. On specific benchmarks (e.g., Prophesee Gen1 dataset), SNNs have even achieved results comparable to or surpassing ANNs, while concurrently demonstrating substantial reductions in energy consumption. This swift improvement in SNN performance, particularly the trend of "disappearing performance differences", signifies a critical transition in SNN technology from purely theoretical research to practical viability. This makes SNNs an increasingly competitive and compelling choice in scenarios with stringent energy efficiency requirements.

The event-driven nature of SNNs makes them highly compatible with specialized neuromorphic chips, such as Intel Loihi[69] and IBM TrueNorth[166], enabling ultra-low power consumption and minimal latency. This capability is indispensable for edge AI applications. The synergistic interplay between hardware and algorithms is a key driving force behind SNNs' performance enhancements and the expansion of their application potential. While SNNs may not yet universally outperform ANNs across all general visual tasks, their advantages are significantly amplified in challenging specific scenarios, including high-speed processing, low-light conditions, and edge deployments. In these contexts, their integration with event cameras and neuromorphic hardware provides a unique and compelling solution for addressing persistent "pain points" in traditional AI systems.

Performance benchmarks for object detection

Significant progress has been made in SNN-based object detection. Spiking CenterNet[78] achieves 2.6% higher mAP than comparable SNNs with better power efficiency. SpikeFPN[79] attains 0.477 mAP on GEN1, demonstrating energy efficiency for automotive event data. SpikiLi[95] achieves CNN-comparable precision with 3ms inference latency for LiDAR-based 3D object detection. The multi-patch localization SNN[122] yields 98.9% accuracy with 0.336 W power and 20FPS for infrared drone detection. The sparse compressed SNN accelerator[101] achieves 26x model size reduction and 1.05 mJ/frame energy efficiency. EMS-YOLO[123] shows ANN-comparable performance with 5.83x less energy and 4 timesteps. Directly trained high-performance SNNs[124] solve gradient vanishing and achieve high accuracy, with spiking RetinaNet reaching 0.296 mAP on MSCOCO. SUHD[97] reduces timesteps by 750x and enhances mAP by 30%, providing ultralow latency. BN-SNN with BIF neurons[87] achieves 0.476 mAP@0.5 on MS-COCO and 0.591 on GEN1 with reduced temporal steps. SpikingYOLOX[77] achieves state-of-the-art performance among SNN-based methods. SFDNet[119] achieves SOTA low-power and robust RGB-event fusion-based object detection. Spike-BRGNet[125] achieves SOTA results on DDD17 and DSEC datasets, outperforming existing SNN methods by +1.57% and +1.91% mIoU respectively, while consuming 17.76x less energy than ANN-based models.

Performance benchmarks for semantic segmentation

In semantic segmentation, SNNs have also made notable advances. EvSegSNN[86] achieves 45.54% MIoU and 89.90% accuracy on the DDD17 dataset with reduced parameters. SpikingEDN[64] achieves 72.57% MIoU on DDD17 and 58.32% on DSEC-Semantic, demonstrating competitive performance with reduced computational demands. SLTNet[144] achieves significant mIoU improvement, 4.58x energy reduction, and 114 FPS for semantic segmentation. The Hybrid SFCNN[147] achieves an mIoU almost 30% higher than existing spiking FCNs on VOC2012. Chakravarty et al.[81] achieve DSC close to 99% on "EM segmentation 2015" and "Data Science Bowl 2018" for image segmentation. Dakic et al.[82] demonstrate similar performance to CNNs while significantly outperforming energy detection methods in spectrum monitoring. Li et al.[89] achieve comparable performance to non-spiking U-Net models, surpassing existing SNN methods, and reduce inference time by approximately 90%. Lei et al.[88] achieve state-of-the-art performance on ADE20K, VOC2012, and CityScapes datasets, highlighting SNN potential for complex segmentation with 5.0x-6.6x efficiency. Sun et al.[145] achieve mIoU scores of 53.7% on ADE20K and 58.06% on pascal context. Li et al.[146] achieve average dice scores of 0.9817 (OD) and 0.9859 (OC) on RIM-ONE-r3, 0.9673 (OD) and 0.9317 (OC) on Drishti-GS, and 0.9687 (OD) and 0.9190 (OC) on REFUGE. Yang et al.[138] achieve high acc (93.87%) and DSC (88.21%) on ISIC2018, demonstrating superior performance and fast processing speed. Ye et al.[80] achieve 43.2% mIoU on PASCAL VOC2012 and 53.4% mIoU on DDD17, with only 2 time steps. Ma et al.[91] achieve 90.4% mIoU and 98.3% PixAcc on diabetic retinal vessel segmentation datasets, demonstrating SOTA energy efficiency. Zhang et al.[104] achieve 94.3% mAP for Sclerotinia detection, with high accuracy and low energy consumption.

The synergistic interplay between hardware and algorithms is a key driving force behind SNNs' performance enhancements and the expansion of their application potential. While SNNs may not yet universally outperform ANNs across all general visual tasks, their advantages are significantly amplified in challenging specific scenarios, including high-speed processing, low-light conditions, and edge deployments. In these contexts, their integration with event cameras and neuromorphic hardware provides a unique and compelling solution for addressing persistent "pain points" in traditional AI systems.

Synergy with optoelectronic sensors and neuromorphic hardware

To emphasize the unique position of SNNs within intelligent opto-electronic systems, this subsection discusses the synergy between SNNs, optoelectronic sensors, and neuromorphic hardware. SNNs are inherently well-suited to process data from event-driven optoelectronic sensors, which generate sparse, asynchronous data streams that mirror the spiking nature of SNNs. This natural compatibility positions SNNs as ideal candidates for low-power, real-time perception in scenarios where traditional frame-based systems struggle.

We explicitly outline the critical functional requirements for SNN-compatible sensors:

1) High temporal resolution and low latency: These sensors must efficiently capture fast-changing dynamic scenes without motion blur, providing data streams that match the ability of SNNs to process information with minimal delay.

2) Event-driven/asynchronous data output: The sensors should fundamentally align with the SNN computational paradigm by only transmitting data (events/spikes) when changes occur, avoiding redundant information transmission.

3) High dynamic range and low power consumption: To meet the demands of edge AI devices, these sensors need to operate effectively in challenging lighting conditions (from dim to bright environments) while consuming minimal power.

These sensor characteristics perfectly complement the design philosophy of neuromorphic hardware (e.g., Intel Loihi), which is optimized for sparse, event-driven computation. The synergy allows for the creation of highly efficient, end-to-end perception-and-computation systems where data is processed directly in the spike domain, from sensor to network, enabling unparalleled energy savings and real-time responsiveness for intelligent opto- electronics.

Key challenges ahead

Despite the remarkable progress, SNNs in image object detection and semantic segmentation still face several critical challenges that need to be addressed for broader adoption.

Training difficulty

The inherent discrete and non-differentiable nature of spike operations makes SNN training substantially more complex than ANNs. Issues like vanishing/exploding gradients persist, particularly in very large and deep SNNs, requiring considerable optimization[81]. The challenges include ensuring high-fidelity information propagation, formulating effective training strategies[89], and managing complex neuronal dynamics and binary activations that lead to performance degradation and non-convergence[88]. SNNs often struggle with training efficiency due to non-differentiable spikes and high memory overhead, hindering deep SNN training[138].

Performance gap and generalization ability

While SNNs have shown excellent results on specific datasets, their generalization ability and absolute performance on larger, more diverse, and complex real-world datasets still need to improve to fully match state-of-the-art ANNs. This disparity currently limits the widespread deployment of SNNs in general-purpose object detection and semantic segmentation tasks. Existing SNN models for image segmentation tend to perform poorly, often underperforming ANNs[88]. Transformer models may cause local information loss, while CNNs struggle with global context, posing challenges for semantic segmentation accuracy[145]. SNNs generally struggle with generalization on small datasets compared to pre-trained models[138].

Hardware support and commercialization

Despite the promising future of neuromorphic hardware, its commercial availability and widespread adoption remain limited. This constraint prevents SNNs from fully realizing their energy efficiency advantages when executed on conventional GPU/CPU platforms, thereby hindering their broader real-world application. Implementing efficient training algorithms for specialized neuromorphic processors remains a key challenge[82]. Deploying complex SNN architectures on neuromorphic chips requires significant computational resources and memory, making it challenging for real-time applications[91]. The synergistic interplay between hardware and algorithms is a key driving force behind SNNs' performance enhancements and the expansion of their application potential.

Information fidelity in spatio-temporal event streams

When SNNs process sparse event data, there is a risk of information loss, especially within discrete binary activations and complex spatio-temporal dynamics[167]. This can adversely affect the model's accuracy and its capacity to capture fine-grained details necessary for pixel-level tasks like semantic segmentation. Creating a universal encoder-decoder for SNNs is difficult for complex and RGB datasets[81]. Inconsistent spike firing rates in skip connections due to data distribution variability can lead to information loss[89]. Spike degradation phenomenon occurs in Mask2Former's deformable attention and mask embedding layers, leading to information loss and reduced firing rates[88]. Traditional CNNs struggle to capture global features due to kernel size limitations, and Transformer models may lose local information[145].

Long Simulation Timesteps

Many SNNs, especially those relying on rate coding or ANN-to-SNN conversion, require a significant number of simulation timesteps to accumulate sufficient information for accurate inference. This can negate some of the latency advantages and increase computational overhead in practical scenarios[88]. SNNs often require multiple time steps for neuron accumulation before firing, which increases computational delay[138]. ANN-to-SNN conversion requires a large number of time-steps for forward inference, leading to high computational redundancy[80]. Spiking-LSTM necessitates multiple simulation steps to achieve desired spike firing rates, increasing time, training, and application costs[104].

These challenges are not isolated but rather deeply interconnected. Training difficulty directly impacts SNNs' performance and generalization capabilities. Concurrently, limited neuromorphic hardware support restricts the full exploitation of SNNs' energy efficiency benefits, which, in turn, constrains their practical adoption and the generation of large-scale datasets, further impeding generalization. To achieve a breakthrough in SNN technology, simultaneous advancements across multiple interconnected layers are imperative: algorithms (e.g., training algorithms, neuron models), models (e.g., architectural design), and hardware (e.g., co-design, commercialization). Addressing these systemic bottlenecks comprehensively is critical for SNNs' maturation.

Future research directions

Future research in Spiking Neural Networks for image object detection and semantic segmentation will predominantly concentrate on several pivotal areas to overcome existing challenges and fully unlock their transformative potential. Additionally, beyond vision tasks, SNNs demonstrate significant potential in other domains such as robotics (for rapid response and control), autonomous driving (for robust sensor fusion and decision-making), and biomedical signal processing (e.g., for brain-computer interfaces), providing relevant citations for further exploration. This versatility highlights the broad applicability of SNN technology.

Efficient and Scalable Training Algorithms

Continued efforts will focus on developing novel learning rules and optimization strategies to tackle inherent training difficulties, aiming for stable and highly efficient training of large-scale, deep SNNs. This includes refining surrogate gradient methods (e.g., exploring adaptive or learnable surrogate functions) and further optimizing Threshold-Dependent Batch Normalization. Additionally, investigating meta-learning or neural architecture search (NAS) specifically for SNNs could automate and optimize training processes. Online optimization of learnable thresholds for improved hardware compatibility and superior performance is a promising avenue[85]. Future research should also focus on universal encoder-decoder frameworks for SNNs capable of converting any RGB image into spiking domain representations with high fidelity[81]. Extending SNNs to dense prediction tasks with sophisticated designs focusing on reducing information loss[88], and deploying Spiking-UNet on neuromorphic chips for image super-resolution[89] are also crucial. NSNPFormer can be extended to other vision tasks and integrated with alternative Transformer backbones to enhance local information capture[145]. Further promotion of SNP application in attention mechanisms and pre-training models[138], and exploring deployment on neuromorphic chips for Spiking-NSNet and Spiking-SSegNet[80] are also vital. Extending Spike-BRGNet to other fields like simultaneous localization and mapping, and flow estimation will broaden SNN applications[125].

Novel Neuron Models and Architectures for Optical Perception

Research will persist in exploring and developing more advanced neuron models, such as dynamic threshold LIF (DT-LIF) and parameterized LIF (PLIF) neurons. These models enhance adaptability by allowing membrane dynamics and thresholds to be learned, significantly improving inference speed and accuracy, particularly in deeper SNNs. Concurrently, new SNN architectures will be designed to better capture and process spatio-temporal information, including further optimizing convolutional SNNs (S-CNNs) to overcome their local and single-scale feature limitations, and exploring more effective multi-scale feature fusion mechanisms. Innovations in recurrent SNNs for temporal reasoning and graph SNNs for relational learning are also promising avenues. Specific directions include dynamic threshold LIF neurons and novel fusion architectures for multi-modal optical data (e.g., event streams + hyperspectral imaging).

Deepened Hybrid Paradigm Integration

Further research will explore the advanced integration of SNNs with traditional ANNs and state-of-the-art models like Transformers. This approach aims to synergistically combine SNNs' energy efficiency and temporal processing capabilities (often for low-level feature extraction from event data) with ANNs' high precision and robust training (for high-level tasks like classification and regression), thereby achieving an optimal balance of performance and energy efficiency. This might involve developing more sophisticated cross-modal fusion techniques (e.g., event-frame fusion) and dynamic switching mechanisms between SNN and ANN components based on task complexity or input characteristics.

Hardware-algorithm co-design and commercialization for intelligent optoelectronics

As neuromorphic chips continue to mature, the synergistic co-design of SNN algorithms and specialized hardware will become increasingly critical. This aims to fully exploit SNNs' low power consumption and real-time processing potential on edge devices. Collaborative efforts between academia and industry will focus on accelerating the commercialization and widespread adoption of neuromorphic hardware, making it more accessible for practical applications. Research into hardware-aware neural architecture search and quantization for SNNs will also be vital. This also discusses specific pathways for combining algorithmic optimizations (like quantization-aware training) with the design of neuromorphic photonic chips to accelerate commercialization.

Optical computing for snns (neuromorphic photonics)

A burgeoning frontier in SNN research involves leveraging optical computing for neuromorphic systems. Optical SNNs (OSNNs) exploit the speed of light and the inherent parallelism of photonic integrated circuits to potentially overcome the bandwidth and energy consumption limitations of electronic systems[168]. By encoding and transmitting spikes as optical pulses, OSNNs offer ultra-high speed, low power consumption, and increased connectivity, which could lead to unprecedented computational densities[169]. This direction aims to realize SNNs on neuromorphic photonic platforms, where neurons and synapses are implemented using optical components like vertical-cavity surface emitting lasers (VCSELs) or degenerate optical parametric oscillators (DOPOs)[170]. Challenges include robust optical neuron activation functions, efficient light-matter interaction for synaptic weights, and integration with existing electronic interfaces[171–172]. However, the potential for massively parallel, high-throughput, and energy-efficient SNNs that bypass electrical bottlenecks makes neuromorphic photonics a highly promising avenue for future SNN acceleration and deployment[173–174].

Benchmark datasets and standardization

The establishment of more comprehensive, diverse, and challenging neuromorphic datasets, alongside unified evaluation metrics and standards, is essential. This will foster fair comparisons among different SNN models and algorithms, accelerating progress across the field. Developing benchmarks that specifically emphasize real-time performance, energy efficiency, and robustness to noisy or incomplete event data will be particularly valuable[175].

Robustness, interpretability, and explainability

As SNNs move toward safety-critical applications like autonomous driving, enhancing their robustness to adversarial attacks[176], improving their interpretability (understanding neuron behavior and spike patterns)[177–179], and providing explainable decisions will become paramount research areas.

Biological plausibility and scalability

Striking a balance between biological realism and computational scalability remains a challenge. Future research might explore incorporating more complex biological mechanisms (e.g., dendritic computation[180], neuromodulation[181]) while ensuring the models remain scalable for large-scale real-world problems[182].

These prospective research directions are not isolated but intricately interdependent. For instance, innovations in novel neuron models and architectural designs lay the groundwork for more efficient training algorithms. Concurrently, the integration of hybrid paradigms capitalizes on the strengths of existing ANNs to accelerate SNN adoption. Hardware-algorithm co-design is fundamental to realizing SNNs' ultimate potential, while robust benchmark datasets are indispensable for advancing all research fronts. The long-term evolution of SNNs is envisioned as a continuous, iterative, and convergent process. SNNs are unlikely to entirely supplant ANNs but are poised to deliver optimal solutions in specific, niche application scenarios, fostering a complementary coexistence within the broader AI ecosystem. The ultimate aspiration is to construct AI systems that are more aligned with biological intelligence, exceptionally efficient, and remarkably versatile.

References

1. Voulodimos A, Doulamis N, Doulamis A et al. Deep learning for computer vision: a brief review. *Comput Intell Neurosci* **2018**, 7068349 (2018).
2. Mahadevkar SV, Khemani B, Patil S et al. A review on machine learning styles in computer vision-techniques and future directions. *IEEE Access* **10**, 107293–107329 (2022).
3. Zhao X, Wang LM, Zhang YF et al. A review of convolutional neural networks in computer vision. *Artif Intell Rev* **57**, 99 (2024).
4. Koroteev MV. BERT: a review of applications in natural language processing and understanding. arXiv: 2103.11943, 2021. https://arxiv.org/abs/2103.11943.
5. Alawida M, Mejri S, Mehmood A et al. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. *Information* **14**, 462 (2023).
6. Wu LF, Chen Y, Shen K et al. Graph neural networks for natural language processing: a survey. *Found Trends Mach Learn* **16**, 119–328 (2023).
7. Treviso M, Lee JU, Ji TC et al. Efficient methods for natural language processing: a survey. *Trans Assoc Comput Ling* **11**, 826–860 (2023).
8. Hinton G, Deng L, Yu D et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag* **29**, 82–97 (2012).
9. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* 1097–1105 (Curran Associates Inc. , 2012).
10. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*

6645–6649 (IEEE, 2013).
http://doi.org/10.1109/ICASSP.2013.6638947.

11. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* **521**, 436–444 (2015).
12. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 3645–3650 (Association for Computational Linguistics, 2019). http://doi.org/10.18653/v1/P19-1355.
13. Schwartz R, Dodge J, Smith NA et al. Green AI. *Commun ACM* **63**, 54–63 (2020).
14. Patterson D, Gonzalez J, Le Q et al. Carbon emissions and large neural network training. arXiv: 2104.10350, 2021. https://arxiv.org/abs/2104.10350.
15. Kuutti S, Bowden R, Jin YC et al. A survey of deep learning applications to autonomous vehicle control. *IEEE Trans Intell Transp Syst* **22**, 712–733 (2021).
16. Chen C, Wang CY, Liu B et al. Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. *IEEE Trans Intell Transp Syst* **24**, 13023–13034 (2023).
17. Katare D, Perino D, Nurmi J et al. A survey on approximate edge AI for energy efficient autonomous driving services. *IEEE Commun Surv Tutorials* **25**, 2714–2754 (2023).
18. Bai LY, Cao JN, Zhang MJ et al. Collaborative edge intelligence for autonomous vehicles: opportunities and challenges. *IEEE Network* **39**, 52–60 (2025).
19. Liu J, Xiang JJ, Jin YJ et al. Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: a survey. *Remote Sens* **13**, 4387 (2021).
20. McEnroe P, Wang S, Liyanage M. A survey on the convergence of edge computing and AI for UAVs: opportunities and challenges. *IEEE Internet Things J* **9**, 15435–15459 (2022).
21. Pal OK, Shovon MSH, Mridha MF et al. In-depth review of AI-enabled unmanned aerial vehicles: trends, vision, and challenges. *Discover Artif Intell* **4**, 97 (2024).
22. Tang GY, Ni JJ, Zhao YH et al. A survey of object detection for UAVs based on deep learning. *Remote Sens* **16**, 149 (2024).
23. Pierson HA, Gashler MS. Deep learning in robotics: a review of recent research. *Adv Robot* **31**, 821–835 (2017).
24. Ramasubramanian AK, Mathew R, Preet I et al. Review and application of edge AI solutions for mobile collaborative robotic platforms. *Procedia CIRP* **107**, 1083–1088 (2022).
25. Jin Y, Huang B, Yan YL et al. Edge-based collaborative training system for artificial intelligence-of-things. *IEEE Trans Ind Inf* **18**, 7162–7173 (2022).
26. Yao JC, Zhang SY, Yao Y et al. Edge-cloud polarization and collaboration: a comprehensive survey for AI. *IEEE Trans Knowl Data Eng* **35**, 6866–6886 (2023).
27. Gu HX, Zhao LQ, Han Z et al. AI-enhanced cloud-edge-terminal collaborative network: survey, applications, and future directions. *IEEE Commun Surv Tutorials* **26**, 1322–1385 (2024).
28. Zhou Z, Chen X, Li E et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing. *Proc IEEE* **107**, 1738–1762 (2019).
29. Deng SG, Zhao HL, Fang WJ et al. Edge intelligence: the confluence of edge computing and artificial intelligence. *IEEE Internet Things J* **7**, 7457–7469 (2020).
30. Duan SJ, Wang D, Ren J et al. Distributed artificial intelligence empowered by end-edge-cloud computing: a survey. *IEEE Commun Surv Tutorials* **25**, 591–624 (2023).
31. Shankar V. Edge AI: a comprehensive survey of technologies, applications, and challenges. In *Proceedings of 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)* 1–6 (IEEE, 2024). http://doi.org/10.1109/ACET61898.2024.10730112.
32. Maass W. Networks of spiking neurons: the third generation of neural network models. *Neural Networks* **10**, 1659–1671 (1997).
33. Gerstner W, Kistler WM, Naud R et al. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition* (Cambridge University Press, Cambridge, 2014).
34. Izhikevich EM. Simple model of spiking neurons. *IEEE Trans Neural Networks* **14**, 1569–1572 (2003).
35. Yu Q, Tang HJ, Tan KC et al. A brain-inspired spiking neural network model with temporal encoding and learning. *Neurocomputing* **138**, 3–13 (2014).
36. Li SX, Zhang ZM, Mao RX et al. A fast and energy-efficient SNN processor with adaptive clock/event-driven computation scheme and online learning. *IEEE Trans Circuits Syst I: Regul Pap* **68**, 1543–1552 (2021).
37. Zhang AG, Li XM, Gao YM et al. Event-driven intrinsic plasticity for spiking convolutional neural networks. *IEEE Trans Neural Networks Learn Syst* **33**, 1986–1995 (2022).
38. Zhang AG, Niu YZ, Gao YM et al. Second-order information bottleneck based spiking neural networks for sEMG recognition. *Inf Sci* **585**, 543–558 (2022).
39. Kim Y, Li YH, Park H et al. Exploring temporal information dynamics in spiking neural networks. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence* 8308–8316 (AAAI, 2023). http://doi.org/10.1609/aaai.v37i7.26002.
40. Zhang AG, Shi JM, Wu JY et al. Low latency and sparse computing spiking neural networks with self-driven adaptive threshold plasticity. *IEEE Trans Neural Networks Learn Syst* **35**, 17177–17188 (2024).
41. Wu KL, E SZ, Yang N et al. A novel approach to enhancing biomedical signal recognition via hybrid high-order information bottleneck driven spiking neural networks. *Neural Networks* **183**, 106976 (2025).
42. Zheng SJ, Qian L, Li PS et al. An introductory review of spiking neural network and artificial neural network: from biological intelligence to artificial intelligence. arXiv: 2204.07519, 2022. https://arxiv.org/abs/2204.07519.
43. Zheng TY, Han LY, Zhang TL. Research advances and new paradigms for biology-inspired spiking neural networks. arXiv: 2408.13996, 2024. https://arxiv.org/abs/2408.13996.
44. Yu W, Yang N, Wang ZJ et al. Fault-tolerant attitude tracking control driven by spiking NNS for unmanned aerial vehicles. *IEEE Trans Neural Networks Learn Syst* **36**, 3773–3785 (2025).
45. Ren SQ, He KM, Girshick R et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell* **39**, 1137–1149 (2017).
46. Redmon J, Divvala S, Girshick R et al. You only look once: unified, real-time object detection. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 779–788 (IEEE, 2016). http://doi.org/10.1109/CVPR.2016.91.
47. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3431–3440 (IEEE, 2015). http://doi.org/10.1109/TPAMI.2016.2572683.
48. Garcia-Garcia A, Orts-Escolano S, Oprea S et al. A review on deep learning techniques applied to semantic segmentation. arXiv: 1704.06857, 2017. https://arxiv.org/abs/1704.06857.
49. Zou ZX, Chen KY, Shi ZW et al. Object detection in 20 years: a survey. *Proc IEEE* **111**, 257–276 (2023).
50. Kaur R, Singh S. A comprehensive review of object detection with deep learning. *Digital Signal Process* **132**, 103812 (2023).
51. Muzammul M, Li X. Comprehensive review of deep learning-based tiny object detection: challenges, strategies, and future directions. *Knowl Inf Syst* **67**, 3825–3913 (2025).
52. Zhou TF, Wang WG, Konukoglu E et al. Rethinking semantic segmentation: a prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2572–2583 (IEEE, 2022). http://doi.org/10.1109/CVPR52688.2022.00261.
53. Mo YJ, Wu Y, Yang XN et al. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **493**, 626–646 (2022).
54. Thisanke H, Deshan C, Chamith K et al. Semantic segmentation using vision transformers: a survey. *Eng Appl Artif Intell* **126**,

106669 (2023).

55. Dayal U. Understanding semantic segmentation: key challenges, techniques, and real-world applications. (2025). https://www.digitaldividedata.com/blog/semantic-segmentation-key-challenges-techniques-and-real-world-applications.
56. Li YH, Deng SK, Dong X et al. A free lunch from ANN: towards efficient, accurate spiking neural networks calibration. In *Proceedings of the 38th International Conference on Machine Learning* 6316–6325 (PMLR, 2021).
57. Gehrig D, Loquercio A, Derpanis KG et al. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 5632–5642 (IEEE, 2019). http://doi.org/10.1109/ICCV.2019.00573.
58. Gallego G, Delbruck T, Orchard G et al. Event-based vision: a survey. *IEEE Trans Pattern Anal Mach Intell* **44**, 154–180 (2022).
59. Shariff W, Dilmaghani MS, Kielty P et al. Event cameras in automotive sensing: a review. *IEEE Access* **12**, 51275–51306 (2024).
60. Chakravarthi B, Verma AA, Daniilidis K et al. Recent event camera innovations: a survey. In *Proceedings of European Conference on Computer Vision* 342–376 (Springer, 2024). http://doi.org/10.1007/978-3-031-92460-6_21.
61. Zhang AG, Song YD. Spiking neural networks in intelligent control systems: a perspective. *Sci China Inf Sci* **67**, 176201 (2024).
62. Barchid S, Mennesson J, Djéraba C. Bina-rep event frames: a simple and effective representation for event-based cameras. In *Proceedings of 2022 IEEE International Conference on Image Processing (ICIP)* 3998–4002 (IEEE, 2022). http://doi.org/10.1109/ICIP46576.2022.9898061.
63. Barchid S, Mennesson J, Djéraba C. Exploring joint embedding architectures and data augmentations for self-supervised representation learning in event-based vision. In *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 3903–3912 (IEEE, 2023). http://doi.org/10.1109/CVPRW59228.2023.00405.
64. Zhang R, Leng LZW, Che KW et al. Accurate and efficient event-based semantic segmentation using adaptive spiking encoder-decoder network. *IEEE Trans Neural Networks Learn Syst* **36**, 9326–9340 (2025).
65. Zhang AG, Han Y, Niu YZ et al. Self-evolutionary neuron model for fast-response spiking neural networks. *IEEE Trans Cogn Dev Syst* **14**, 1766–1777 (2022).
66. Barchid S, Mennesson J, Eshraghian J et al. Spiking neural networks for frame-based and event-based single object localization. *Neurocomputing* **559**, 126805 (2023).
67. Zhang AG, Gao YM, Niu YZ et al. Intrinsic plasticity for online unsupervised learning based on soft-reset spiking neuron model. *IEEE Trans Cogn Dev Syst* **15**, 337–347 (2023).
68. Zhou Y, Li XY, Wu XY et al. Object detection method with spiking neural network based on DT-LIF neuron and SSD. *J Electron Inf Technol* **45**, 2722–2730 (2023).
69. Davies M, Srinivasa N, Lin TH et al. Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99 (2018).
70. Lichtsteiner P, Posch C, Delbruck T. A 128× 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J Solid-State Circuits* **43**, 566–576 (2008).
71. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* **117**, 500–544 (1952).
72. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).
73. Furber SB, Galluppi F, Temple S et al. The spinnaker project. *Proc IEEE* **102**, 652–665 (2014).
74. Brette R, Gerstner W. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J Neurophysiol* **94**, 3637–3642 (2005).
75. Strukov D, Indiveri G, Grollier J et al. Building brain-inspired computing. *Nat Commun* **10**, 4838 (2019).
76. Kim S, Park S, Na B et al. Spiking-YOLO: spiking neural network for energy-efficient object detection. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence* 11270–11277 (AAAI, 2020). http://doi.org/10.1609/aaai.v34i07.6787.
77. Miao W, Shen JR, Xu Q et al. SpikingYOLOX: improved YOLOX object detection with fast Fourier convolution and spiking neural networks. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence* 1465–1473 (AAAI, 2025). http://doi.org/10.1609/aaai.v39i2.32137.
78. Bodden L, Ha DB, Schwaiger F et al. Spiking CenterNet: a distillation-boosted spiking neural network for object detection. In *Proceedings of 2024 International Joint Conference on Neural Networks (IJCNN)* 1–9 (IEEE, 2024). http://doi.org/10.1109/IJCNN60899.2024.10650418.
79. Zhang H, Li YC, Leng LZW et al. Automotive object detection via learning sparse events by spiking neurons. *IEEE Trans Cogn Dev Syst* **16**, 2110–2124 (2024).
80. Ye WJ, Chen SZ, Liu HX et al. The architecture design and training optimization of spiking neural network with low-latency and high-performance for classification and segmentation. *Neural Networks* **191**, 107790 (2025).
81. Chakravarty S, Tanveer MH, Voicu RC et al. Backpropagation techniques in SNN and application in image segmentation. In *Proceedings of SoutheastCon 2024* 1475–1481 (IEEE, 2024). http://doi.org/10.1109/SoutheastCon52093.2024.10500286.
82. Dakic K, Homssi BA, Al-Hourani A. Spiking-UNet: spiking neural networks for spectrum occupancy monitoring. In *Proceedings of 2024 IEEE Wireless Communications and Networking Conference (WCNC)* 1–6 (IEEE, 2024). http://doi.org/10.1109/WCNC57260.2024.10571312.
83. Fang W, Yu ZF, Chen YQ et al. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 2641–2651 (IEEE, 2021). http://doi.org/10.1109/ICCV48922.2021.00266.
84. Courtois J, Novac PE, Lemaire E et al. Embedded event based object detection with spiking neural network. In *Proceedings of 2024 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2024). http://doi.org/10.1109/IJCNN60899.2024.10649943.
85. Hasssan A, Meng J, Seo JS. Spiking neural network with learnable threshold for event-based classification and object detection. In *Proceedings of 2024 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2024). http://doi.org/10.1109/IJCNN60899.2024.10650320.
86. Hareb D, Martinet J. EvSegSNN: neuromorphic semantic segmentation for event data. In *Proceedings of 2024 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2024). http://doi.org/10.1109/IJCNN60899.2024.10650811.
87. Yasir SM, Kim H. BN-SNN: spiking neural networks with bistable neurons for object detection. *PLoS One* **20**, e0327513 (2025).
88. Lei ZX, Yao M, Hu JK et al. Spike2Former: efficient spiking transformer for high-performance image segmentation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)* 1364–1372 (AAAI, 2025). http://doi.org/10.1609/aaai.v39i2.32126.
89. Li HB, Zhang YY, Xiong ZW et al. Deep multi-threshold spiking-UNet for image processing. *Neurocomputing* **586**, 127653 (2024).
90. Ding JC, Dong B, Heide F et al. Biologically inspired dynamic thresholds for spiking neural networks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* 441 (Curran Associates Inc. , 2022).
91. Ma YQ, Wang HM, Shen HC et al. Analog spiking U-Net integrating CBAM&ViT for medical image segmentation. *Neural Networks* **181**, 106765 (2025).
92. Kachole S, Sajwani H, Naeini FB et al. Asynchronous bioplausible neuron for spiking neural networks for event-based vision. In *Proceedings of the 18th European Conference on Computer Vision* 399–415 (Springer, 2024).

http://doi.org/10.1007/978-3-031-73039-9_23.

93. Thorpe S, Delorme A, Van Rullen R. Spike-based strategies for rapid processing. *Neural Networks* **14**, 715–725 (2001).
94. Diehl PU, Cook M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front Comput Neurosci* **9**, 149773 (2015).
95. Mohapatra S, Mesquida T, Hodaei M et al. SpikiLi: a spiking simulation of LiDAR based real-time object detection for autonomous driving. In *Proceedings of 2022 8th International Conference on Event-Based Control, Communication, and Signal Processing (EBCCSP)* 1–5 (IEEE, 2022). http://doi.org/10.1109/EBCCSP56922.2022.9845647.
96. Li D, Jin JD, Zhang YH et al. Semantic-aware frame-event fusion based pattern recognition via large vision-language models. *Pattern Recognit* **158**, 111080 (2025).
97. Qu JY, Gao ZY, Zhang TL et al. Spiking neural network for ultralow-latency and high-accurate object detection. *IEEE Trans Neural Networks Learn Syst* **36**, 4934–4946 (2025).
98. Esser SK, Merolla PA, Arthur JV et al. Convolutional networks for fast, energy-efficient neuromorphic computing. *Proc Natl Acad Sci USA* **113**, 11441–11446 (2016).
99. Hunsberger E, Eliasmith C. Training spiking deep networks for neuromorphic hardware. arXiv: 1510.08829, 2016. https://arxiv.org/abs/1611.05141.
100. Balaji A, Catthoor F, Das A et al. Mapping spiking neural networks to neuromorphic hardware. *IEEE Trans Very Large Scale Integr (VLSI) Syst* **28**, 76–86 (2020).
101. Lien HH, Chang TS. Sparse compressed spiking neural network accelerator for object detection. *IEEE Trans Circuits Syst I: Regul Pap* **69**, 2060–2069 (2022).
102. Liu BC, Yu Q, Gao JW et al. Spiking neuron networks based energy-efficient object detection for mobile robot. In *Proceedings of 2021 China Automation Congress (CAC)* 3224–3229 (IEEE, 2021). http://doi.org/10.1109/CAC53003.2021.9727350.
103. Wang HZ, Liang XB, Zhang T et al. PSSD-transformer: powerful sparse spike-driven transformer for image semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia* 758–767 (Association for Computing Machinery, 2024). http://doi.org/10.1145/3664647.3680870.
104. Zhang J, Zhao YR, Yan J et al. Spiking-LSTM: a novel hyperspectral image segmentation network for Sclerotinia detection. *Comput Electron Agric* **226**, 109397 (2024).
105. Rebecq H, Ranftl R, Koltun V et al. High speed and high dynamic range video with an event camera. *IEEE Trans Pattern Anal Mach Intell* **43**, 1964–1980 (2021).
106. López-Randulfe J, Duswald T, Bing ZS et al. Spiking neural network for Fourier transform and object detection for automotive radar. *Front Neurorobot* **15**, 688344 (2021).
107. Cabarle FGC. Thinking about spiking neural p systems: some theories, tools, and research topics. *J Membr Comput* **6**, 148–167 (2024).
108. Zandron C. An overview on applications of spiking neural networks and spiking neural p systems. In López MDJ, Vaszil G. *Languages of Cooperation and Communication* 267–278 (Springer, Cham, 2025).
109. Schuman CD et al. A survey of neuromorphic computing and neural networks in hardware. arXiv: 1705.06963, 2017. https://arxiv.org/abs/1705.06963.
110. Shrestha A, Fang HW, Mei ZD et al. A survey on neuromorphic computing: models and hardware. *IEEE Circuits Syst Mag* **22**, 6–35 (2022).
111. Thakur D, Guzzo A, Fortino G. Hardware-algorithm co-design of energy efficient federated learning in quantized neural network. *Internet Things* **26**, 101223 (2024).
112. Bohte SM, Kok JN, La Poutré H. Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing* **48**, 17–37 (2002).
113. Diehl PU, Neil D, Binas J et al. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *Proceedings of 2015 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2015). http://doi.org/10.1109/IJCNN.2015.7280696.
114. Meftah B, Lezoray O, Benyettou A. Segmentation and edge detection based on spiking neural network model. *Neural Process Lett* **32**, 131–146 (2010).
115. Cao YQ, Chen Y, Khosla D. Spiking deep convolutional neural networks for energy-efficient object recognition. *Int J Comput Vis* **113**, 54–66 (2015).
116. Bi T, Rong HN, Zhou ZZ et al. Spiking neural network based on YOLO-C3 for object detection. In *Proceedings of 2024 4th International Symposium on Artificial Intelligence and Intelligent Manufacturing (AIIM)* 608–612 (IEEE, 2024). http://doi.org/10.1109/AIIM64537.2024.10934233.
117. Fu Q, Dong HB. Spiking neural network based on multi-scale saliency fusion for breast cancer detection. *Entropy* **24**, 1543 (2022).
118. Fan YM, Zhang W, Liu CS et al. SFOD: spiking fusion object detector. In *Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 17191–17200 (IEEE, 2024). http://doi.org/10.1109/CVPR52733.2024.01627.
119. Fan LW, Yang JJ, Wang L et al. Efficient spiking neural network for RGB-event fusion-based object detection. *Electronics* **14**, 1105 (2025).
120. Fan YM, Liu CS, Li MY et al. SpikeDet: better firing patterns for accurate and energy-efficient object detection with spiking neuron networks. arXiv: 2501.15151, 2025. https://arxiv.org/abs/2501.15151.
121. Bulzomi H, Gruel A, Martinet J et al. Object detection for embedded systems using tiny spiking neural networks: filtering noise through visual attention. In *Proceedings of 2023 18th International Conference on Machine Vision and Applications (MVA)* 1–5 (IEEE, 2023). http://doi.org/10.23919/MVA57639.2023.10215590.
122. Feng S, Cao J, Zhang L et al. Multi-patch localization spiking neural network for object detection. In *Proceedings of 2022 IEEE 16th International Conference on Solid-State & Integrated Circuit Technology (ICSICT)* 1–4 (IEEE, 2022). http://doi.org/10.1109/ICSICT55466.2022.9963459.
123. Su QY, Chou YH, Hu YF et al. Deep directly-trained spiking neural networks for object detection. In *Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 6532–6542 (IEEE, 2023). http://doi.org/10.1109/ICCV51070.2023.00603.
124. Zhang H, Li Y, He B et al. Direct training high-performance spiking neural networks for object recognition and detection. *Front Neurosci* **17**, 1229951 (2023).
125. Long XL, Zhu XX, Guo FM et al. Spike-BRGNet: efficient and accurate event-based semantic segmentation with boundary region-guided spiking neural networks. *IEEE Trans Circuits Syst Video Technol* **35**, 2712–2724 (2025).
126. Carion N, Massa F, Synnaeve G et al. End-to-end object detection with transformers. In *Proceedings of the 16th European Conference on Computer Vision* 213–229 (Springer, 2020). http://doi.org/10.1007/978-3-030-58452-8_13.
127. Zhou ZK, Zhu YS, He C et al. Spikformer: when spiking neural network meets transformer. In *Proceedings of the Eleventh International Conference on Learning Representations* (ICLR, 2023).
128. Datta G, Liu ZY, Li AN et al. Spiking neural networks with dynamic time steps for vision transformers. arXiv: 2311.16456, 2023. https://arxiv.org/abs/2311.16456.
129. Gong L, Dong H, Zhang XY et al. Spiking ViT: spiking neural networks with transformer-attention for steel surface defect classification. *J Electron Imaging* **33**, 033001 (2024).
130. Li YD, Lei YL, Yang X. Spikeformer: training high-performance spiking neural network with transformer. *Neurocomputing* **574**, 127279 (2024).
131. Zhou ZK, Che KW, Fang W et al. Spikformer V2: join the high accuracy club on ImageNet with an SNN ticket. arXiv: 2401.02020, 2024. https://arxiv.org/abs/2401.02020.

132. Wang S et al. Spiking vision transformer with saccadic attention. In *Proceedings of the 13th International Conference on Learning Representation* (ICLR, 2025).
133. Yao M, Qiu XR, Hu TX et al. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Trans Pattern Anal Mach Intell* **47**, 2973–2990 (2025).
134. Yu LX, Chen HQ, Wang ZM et al. SpikingViT: a multiscale spiking vision transformer model for event-based object detection. *IEEE Trans Cogn Dev Syst* **17**, 130–146 (2025).
135. Woo S, Park J, Lee JY et al. CBAM: convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision* 3–19 (Springer, 2018). http://doi.org/10.1007/978-3-030-01234-2_1.
136. Lin TY, Dollár P, Girshick R et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 936–944 (IEEE, 2017). http://doi.org/10.1109/CVPR.2017.106.
137. Liu Z, Lin YT, Cao Y et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 9992–10002 (2021). http://doi.org/10.1109/ICCV48922.2021.00986.
138. Yang B, Zhang RM, Peng H et al. SLP-Net: an efficient lightweight network for segmentation of skin lesions. *Biomed Signal Process Control* **101**, 107242 (2025).
139. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* **39**, 2481–2495 (2017).
140. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* 234–241 (Springer, 2015). http://doi.org/10.1007/978-3-319-24574-4_28.
141. Chen LC, Papandreou G, Schroff F et al. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017. https://arxiv.org/abs/1706.05587.
142. Zhao HS, Shi JP, Qi XJ et al. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6230–6239 (IEEE, 2017). http://doi.org/10.1109/CVPR.2017.660.
143. Isensee F, Jaeger PF, Kohl SAA et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203–211 (2021).
144. Long XL, Zhu XX, Guo FM et al. SLTNet: efficient event-based semantic segmentation with spike-driven lightweight transformer-based networks. arXiv: 2412.12843, 2024. https://arxiv.org/pdf/2412.12843.
145. Sun SY, Yang WQ, Peng H et al. A semantic segmentation method integrated convolutional nonlinear spiking neural model with transformer. *Comput Vision Image Understanding* **249**, 104196 (2024).
146. Li W, Xia MC, Peng H et al. ODCS-NSNP: optic disc and cup segmentation using deep networks enhanced by nonlinear spiking neural P systems. *Biomed Signal Process Control* **108**, 107935 (2025).
147. Zhang T, Xiang SY, Liu WZ et al. Hybrid spiking fully convolutional neural network for semantic segmentation. *Electronics* **12**, 3565 (2023).
148. Li TP, Cui ZT, Han Y et al. Enhanced multi-scale networks for semantic segmentation. *Complex Intell Syst* **10**, 2557–2568 (2024).
149. Li HB, Peng YS, Yuan JH et al. Efficient event-based semantic segmentation via exploiting frame-event fusion: a hybrid neural network approach. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)* 18296–18304 (AAAI, 2025). http://doi.org/10.1609/aaai.v39i17.34013.
150. Neftci EO, Mostafa H, Zenke F. Surrogate gradient learning in spiking neural networks: bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Process Mag* **36**, 51–63 (2019).
151. Li YH, Guo YF, Zhang SH et al. Differentiable spike: rethinking gradient-descent for training spiking neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems* 1794 (Curran Associates Inc. , 2021).
152. Huh D, Sejnowski TJ. Gradient descent for spiking neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* 1440–1450 (Curran Associates Inc. , 2018).
153. Werbos PJ. Backpropagation through time: what it does and how to do it. *Proc IEEE* **78**, 1550–1560 (1990).
154. Lillicrap TP, Santoro A. Backpropagation through time and the brain. *Curr Opin Neurobiol* **55**, 82–89 (2019).
155. Dampfhoffer M, Mesquida T, Valentian A et al. Backpropagation-based learning techniques for deep spiking neural networks: a survey. *IEEE Trans Neural Networks Learn Syst* **35**, 11906–11921 (2024).
156. Wu YJ, Deng L, Li GQ et al. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Front Neurosci* **12**, 331 (2018).
157. Zhang WR, Li P. Temporal spike sequence learning via backpropagation for deep spiking neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* 1008 (Curran Associates Inc. , 2020).
158. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning* 448–456 (JMLR. org, 2015).
159. Zheng HL, Wu YJ, Deng L et al. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence* 11062–11070 (AAAI, 2021). http://doi.org/10.1609/aaai.v35i12.17320.
160. Diehl PU, Zarrella G, Cassidy A et al. Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In *Proceedings of 2016 IEEE International Conference on Rebooting Computing (ICRC)* 1–8 (IEEE, 2016). http://doi.org/10.1109/ICRC.2016.7738691.
161. Li Y, He X, Dong YT et al. Spike calibration: fast and accurate conversion of spiking neural network for object detection and segmentation. arXiv: 2207.02702, 2022. https://arxiv.org/abs/2207.02702.
162. Neftci EO, Augustine C, Paul S et al. Event-driven random backpropagation: enabling neuromorphic deep learning machines. *Front Neurosci* **11**, 324 (2017).
163. Zhu YY, Yu ZF, Fang W et al. Training spiking neural networks with event-driven backpropagation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* 2213 (Curran Associates Inc. , 2022).
164. Wei WJ, Zhang ML, Qu H et al. Temporal-coded spiking neural networks with dynamic firing threshold: learning with event-driven backpropagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 10518–10528 (IEEE, 2023). http://doi.org/10.1109/ICCV51070.2023.00968.
165. Wei WJ, Zhang ML, Zhang JL et al. Event-driven learning for spiking neural networks. arXiv: 2403.00270, 2024. https://arxiv.org/abs/2403.00270.
166. Akopyan F, Sawada J, Cassidy A et al. TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **34**, 1537–1557 (2015).
167. Song TY, Jin GY, Li PP et al. Learning a spiking neural network for efficient image deraining. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence* 139 (Curran Associates, 2024). http://doi.org/10.24963/ijcai.2024/139.
168. Feldmann J, Youngblood N, Wright CD et al. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
169. Xu YF, Tang GZ, Yousefzadeh A et al. Event-based optical flow on neuromorphic processor: ANN vs. SNN comparison based on

activation sparsification. *Neural Networks* **188**, 107447 (2025).

170. Xiang SY, Ren ZX, Song ZW et al. Computing primitive of fully VCSEL-based all-optical spiking neural network for supervised learning and pattern classification. *IEEE Trans Neural Networks Learn Syst* **32**, 2494–2505 (2021).
171. Xiang SY, Han YN, Song ZW et al. A review: photonics devices, architectures, and algorithms for optical neural computing. *J Semicond* **42**, 023105 (2021).
172. Kutluyarov RV, Zakoyan AG, Voronkov GS et al. Neuromorphic photonics circuits: contemporary review. *Nanomaterials* **13**, 3139 (2023).
173. Srouji LE, Lee YJ, On MB et al. Scalable nanophotonic-electronic spiking neural networks. *IEEE J Select Topics Quantum Electron* **29**, 6000113 (2023).
174. Brunner D, Shastri BJ, Al Qadasi MA et al. Roadmap on neuromorphic photonics. arXiv: 2501.07917, 2025. https://arxiv.org/abs/2501.07917.
175. Lin JQ, Lu S, Bal M et al. Benchmarking spiking neural network learning methods with varying locality. *IEEE Access* **13**, 113606–113617 (2025).
176. Chen WR, Xu Q. Robust and efficient adversarial defense in SNNs via image purification and joint detection. In *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing* 1–5 (IEEE, 2025). http://doi.org/10.1109/ICASSP49660.2025.10888581.
177. Zhang BX, Xu Z, Tao K. Enhancing generalization of spiking neural networks through temporal regularization. arXiv: 2506.19256, 2025. https://arxiv.org/abs/2506.19256.
178. Wetzel SJ, Ha S, Iten R et al. Interpretable machine learning in physics: a review. arXiv: 2503.23616, 2025. https://arxiv.org/abs/2503.23616.
179. Mersha M, Lam K, Wood J et al. Explainable artificial intelligence: a survey of needs, techniques, applications, and future direction. *Neurocomputing* **599**, 128111 (2024).
180. Capone C, Lupo C, Muratore P et al. Beyond spiking networks: the computational advantages of dendritic amplification and input segregation. *Proc Natl Acad Sci USA* **120**, e2220743120 (2023).
181. AlKilany A, Goodman DFM. Neuromodulation enhances dynamic sensory processing in spiking neural network models. *bioRxiv* (2025).
182. Xiao Y, Liu YZ, Zhang BH et al. Bio-plausible reconfigurable spiking neuron for neuromorphic computing. *Sci Adv* **11**, eadr6733 (2025).

Acknowledgements

We are grateful for financial supports from National Natural Science Foundation of China under Grant No. 62306001, and the Natural Science Foundation of Fujian Province, China under Grant No. 2025J01290.

Author contributions

Anguo Zhang, Hongwei Cao, and Yongduan Song conceived of the idea and designed the review. Anguo Zhang, Hongwei Cao, and Na Shan led the manuscript writing and review, with Jiaqi Wang focusing on the theoretical analysis generation and Mingbo Pu contributing to the application aspects. All authors participated in the review and discussion of the manuscript.

Competing interests

The authors declare no competing financial interests.