Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078314

ARTICLE

DGRDet: Dynamic Gaussian Receptive Field Encoding-Based Spiking Neural Networks for Remote Sensing Object Detection

Li Chen[1], Fan Zhang[2,*], Guangwei Xie[3], Yanzhao Gao[1], Xiaofeng Qi[1] and Mingqian Sun[2]

[1]National Digital Switching System Engineering & Technological R&D Center, Information Engineering University, Zhengzhou, China

[2]School of Computer Science, Fudan University, Shanghai, China

[3]Shanghai HONGZHEN Information Science & Technology Corporation, Shanghai, China

*Corresponding Author: Fan Zhang. Email: zhangfanryan@163.com

Received: 29 December 2025; Accepted: 16 April 2026; Published: 15 June 2026

ABSTRACT: Remote sensing object detection aims to identify and localize specific targets in satellite or aerial imagery. Spiking Neural Networks (SNNs), benefiting from their implicit feedback-based and event-driven brain-inspired dynamics, offer a promising solution to alleviate the high energy consumption of conventional ANN-based detection models. However, existing SNN-based approaches for remote sensing object detection—particularly for small, arbitrarily rotated objects—are still in their infancy and suffer from a substantial performance gap compared with ANN counterparts. In this work, we draw inspiration from the hierarchical sparse perception mechanisms of biological vision and integrate dynamic receptive field modulation into the encoding stage, proposing a high-precision spiking object detection framework tailored for remote sensing image. Specifically, we design a Hierarchical Feedback-based Gaussian Encoding (HFG) scheme, in which the parameters of Gaussian kernels are dynamically adjusted through spike-triggered top-down feedback connections. This mechanism enables the encoding process to adaptively respond to complex geometric variations of remote sensing objects, including rotation and scale changes. Based on the proposed encoding strategy, we develop DGRDet (Dynamic Gaussian Receptive Field Encoding-based Spiking Neural Networks for Remote Sensing Object Detection), a directly trained deep SNN detector for remote sensing image. Extensive evaluations on the large-scale public DOTA dataset demonstrate that DGRDet achieves competitive detection accuracy, outperforming existing SNN-based object detection methods. Moreover, compared with ANN models of comparable detection performance, DGRDet reduces spike activity by 81.31% and requires only 0.12% of the inference energy consumption, achieving a favorable balance between detection accuracy, efficiency, and energy efficiency.

KEYWORDS: Remote sensing image; object detection; spiking neural networks (SNNs); hierarchical sparse; dynamic gaussian encoding

1 Introduction

Remote sensing imagery object detection represents one of the most formidable challenges in computer vision. Precise identification and localization of remote sensing objects are instrumental across diverse sectors [1], including environmental monitoring, military strikes, and the low-altitude economy, providing a critical safeguard for national defense and civil urban planning. However, owing to intrinsic challenges such as viewpoint variations, scaling fluctuations, and occlusions, there remains significant room for optimization in the recognition accuracy of images captured by remote sensing equipment.

In recent years, deep learning algorithms based on Artificial Neural Networks (ANNs) have attracted overwhelming academic attention. These algorithms have achieved remarkable success across various specific computational tasks, often reaching performance levels comparable to human operators. Nevertheless, such superior performance entails immense environmental and energy consumption. Due to the limitations of data throughput between computation and memory, a pronounced energy-efficiency gap persists between the inference mechanisms of conventional deep learning models and the massively parallel and event-driven processing mechanisms of the human brain. Spiking Neural Networks (SNNs) [2], leveraging advantages such as sparse neuronal computation and temporal coding, significantly reduce resource requirements in terms of power, energy, and computation. This remarks SNNs a promising solution for deep learning applications—specifically for computer vision tasks like remote sensing object detection—that must be deployed on resource-constrained edge devices, including unmanned aerial vehicles (UAVs) and handheld Internet-of-Things (IoT) devices.

Despite these advantages, SNNs have not been comprehensively adopted due to the lack of effective training methodologies. The non-differentiability of discrete spikes prevents the direct application of conventional gradient propagation algorithms, while binary discretized spike sequences, to some extent, hinder SNNs from achieving advanced recognition accuracy. As a result, most existing SNN-based algorithms only rival ANNs in simpler tasks such as image classification or handwritten digit recognition [3]; In contrast, SNNs rarely occupy a competitive position in more complex computer vision tasks, including image segmentation and object detection. To strike a balance between performance and efficiency, researchers have proposed ANN-to-SNN conversion methods during this transitional phase [4,5]. These approaches transform pretrained ANN models into high-accuracy SNNs through techniques such as weight normalization and neuron replacement. For example, Spiking-YOLO [6] introduced threshold-balanced neurons and channel-wise normalization to improve conversion convergence accuracy, and was the first to apply SNNs to natural image object detection. Furthermore, Li et al. [7] corrected spike firing rate to further enhance detection accuracy after conversion. However, such conversion-based methods typically require a large number of time steps, resulting in high latency, which—together with residual performance gaps—prevents their deployment on mobile and edge devices.

To address latency and deployment issues, Jin et al. proposed a region-based SNN [8], achieving high accuracy on the VOC dataset [9]. Qu et al. [10] focused on compressing time steps, utilizing the low-timestep SNN model SUHD to attain detection performance comparable to ANNs on the COCO dataset [11]. Regarding direct training methods for SNNs, EMS-YOLO [12] was first applied to object detection in 2023. In 2024, Meta-SpikeFormer [13] further advanced SNN-based object detection accuracy to new heights through three spike-driven self-attention mechanisms of varying complexity. Nevertheless, a significant performance disparity remains between these models and ANNs. Moreover, all of the aforementioned methods are primarily designed for natural images, leaving the more challenging domain of remote sensing imagery largely unexplored. Our objective is to bridge the performance gap between SNN and ANN models within the remote sensing field and demonstrate the low-power advantages of event-driven computation, thereby providing viable solutions for diverse remote sensing scenarios.

In this research, to effectively enhance the performance of Spiking Neural Networks (SNNs) in real-time remote sensing object detection scenarios, we draw inspiration from the sparse encoding and hierarchical information transmission mechanisms of biological vision. We aim to construct a high-accuracy SNN-based detection model tailored for the remote sensing domain, while achieving superior efficiency compared to existing conventional ANN-based detection algorithms. The core idea is to exploit sparsity to focus on salient features, thereby improving the network's representational capacity for input data, and to refine the functionality of each network layer through a hierarchical architecture.

The main contributions of this study are summarized as follows, with an emphasis on the advantages of the proposed DGRDet framework:

(1) Hierarchically feedback-based Gaussian encoding

Through an in-depth analysis of the limitations inherent in current rate-based and temporal coding methods, and by integrating the characteristics of remote sensing objects with biological feedback-stimulation patterns, we establish a hierarchical feedback-based Gaussian encoding method (HFG). From a theoretical perspective, we demonstrate that the proposed Gaussian encoding scheme can effectively adjust both the shape and spatial position of neuronal receptive fields, thereby enhancing the features of oriented remote sensing objects.

(2) A deep SNN model for remote sensing object detection.

We propose DGRDet, an SNN-based deep model specifically designed for remote sensing image object detection. With an extremely short number of time steps, DGRDet achieves competitive performance and efficient detection on the public remote sensing dataset DOTA [14] (mAP: 70.33%, time steps: 4). Driven by the rapidly evolving field of neuromorphic remote sensing, this work represents one of the early explorations into applying Spiking Neural Networks (SNNs) to high-resolution remote sensing object detection.

The remainder of this paper is organized as follows. Section 2 reviews object detection algorithms in remote sensing scenarios and existing mainstream SNN encoding strategies. Section 3 presents the proposed remote sensing object detection framework DGRDet in detail. Section 4 conducts a two-level comparative evaluation between DGRDet and ANN-based remote sensing detectors, ANN-to-SNN conversion methods, and directly trained SNN approaches, along with ablation studies on the proposed HFG encoding scheme to comprehensively assess detection accuracy and inference efficiency. Finally, Section 5 concludes the paper and discusses future study directions.

2 Related Work

2.1 Remote Sensing Object Detection

As a frontier task in the field of remote sensing imagery, remote sensing object detection has become a testing ground for numerous detection algorithms due to its unique challenges, including small and densely distributed objects (e.g., parking lots), arbitrary orientations, large aspect ratios (e.g., bridges and ports), and significant scale variations among objects. Currently, research on high-precision remote sensing object detection primarily focuses on four aspects: detection frameworks, feature refinement, oriented loss function optimization, and scenario-specific object modeling.

In general, existing detection frameworks for remote sensing objects can be categorized into two-stage detectors, one-stage detectors, anchor-free detectors, and detectors based on DETR (DEtection TRansformer) [15]. Taking two-stage detectors—which center on candidate target regions—as an example, region proposal–based methods combined with Feature Pyramid Networks (FPN) [16], such as Faster R-CNN [17], are commonly regarded as benchmark models due to their efficient design and outstanding accuracy. When extended with orientation-aware formulations, these methods are often referred to as Faster R-CNN OBB (Oriented Bounding Box). However, conventional Region Proposal Networks (RPNs) generate only horizontal Regions of Interest (RoIs), leading to feature misalignment between horizontal RoIs and oriented bounding boxes, as illustrated in Fig. 1.

To address this misalignment issue, Ding et al. [18] introduced a lightweight RoI learning module on top of the original RPN, which transforms horizontal RoIs into oriented ones using a small number of converted anchors, thereby improving efficiency. Nevertheless, this approach incurs additional computational overhead due to the increased number of anchors. Subsequently, Xie et al. [19] proposed a simpler architecture

that achieves a favorable balance between accuracy and efficiency by modeling center-point offsets of bounding boxes. Subsequent two-stage detectors have explored improving performance by enhancing feature representations and refining oriented bounding box (OBB) formulations, for example ARC [20], STD [21], and QPDet [22], which have reported consistent gains on standard remote sensing benchmarks. Despite their strong performance advantages in terms of accuracy, two-stage detectors generally suffer from limited detection efficiency. To improve efficiency while maintaining high accuracy, one-stage detectors have been developed with detection speed as a primary objective. R3Det [23] integrates multi-level remote sensing features through feature refinement and achieves rapid accuracy improvements via refined bounding box regression and target center reconstruction. Sun et al. [24] proposed a spatial transformation selection strategy to dynamically assign classification labels, ensuring sufficient positive samples for objects with large aspect ratios.

(a) (b)

Figure 1: Two types of bounding box representations in remote sensing imagery. Horizontal Bounding Boxes (HBB) (**a**) lead to feature misalignment, whereas Oriented Bounding Boxes (OBB) (**b**) avoid feature overlapping.

However, both of the aforementioned detector categories follow the anchor-based paradigm, and thus inherently suffer from the fundamental contradiction of spatial misalignment between horizontal anchors and oriented detection boxes. This limitation has gradually driven research toward anchor-free detection methods. From the perspective of oriented bounding box formulation, anchor-free approaches can be divided into keypoint-based and center-based methods. In 2022, Li et al. [25] proposed Oriented RepPoints, which evaluate and assign keypoints adaptively to measure keypoint quality without introducing additional computational overhead during inference. In 2024, Xie et al. introduced DFDet [26], which significantly improves detection accuracy under the anchor-free paradigm by incorporating contextual information and a penalty–incentive assignment strategy on top of center-based detection. The aforementioned detection models avoid the cumbersome manual design of anchors and exhibit inherent advantages in detection efficiency; however, they remain within the scope of convolutional neural networks. Beyond convolution-based methods, DETR-based detectors [15] have also achieved remarkable success in remote sensing imagery. Under the DETR framework, the integration of anchor information combined with the utilization of Transformer models enables effective rotated remote sensing object detection. In 2024, Zeng et al. [27] proposed ARS-DETR, which employs an aspect-ratio-aware circular smooth label to more reasonably smooth angular representations and introduces a rotatable attention module to alleviate the spatial misalignment between sampling points and regional features to some extent. Nevertheless, despite their promising accuracy, Transformer-based object detection models still suffer from long training times and high computational costs, which remain critical challenges to be addressed.

Notably, all of the aforementioned methods operate within the ANN framework to align and optimize object-level spatial features. This paradigm inherently leads to high energy consumption. Even with the

fastest one-stage detection algorithms, the resulting detection efficiency and power consumption still fall short of the low-latency and lightweight requirements of remote sensing object detection applications. Recently, Spiking Neural Networks (SNNs) [2] have attracted increasing attention due to their sparse information transmission mechanisms, and their inherent low-power advantages make them particularly promising for remote sensing object monitoring and deployment in energy-constrained environments.

To provide a longitudinal perspective on the evolution of object detection, the aforementioned methods can be conceptually taxonomized into two main branches: ANN-based and SNN-based paradigms. Within the ANN branch, the trajectory has evolved from complex two-stage detectors to highly efficient one-stage methods. Concurrently, the SNN branch has progressed from early ANN-to-SNN conversion techniques toward direct SNN training methods, which offer superior energy efficiency and serve as the foundational paradigm for our proposed DGRDet.

2.2 SNN Encoding Schemes

In the context of SNNs, the encoding process serves as the crucial interface that converts continuous static input data into discrete spatio-temporal spike trains for subsequent network processing. According to mainstream research directions in neuroscience, SNN encoding schemes can generally be categorized into Rate coding and Temporal coding. Rate coding represents information by extending the input over time and encoding it using multiple spikes generated within a predefined temporal window. The firing rate v is commonly used to characterize this process, which can be defined as:

$$v = \frac{N_{spike}}{T}, \tag{1}$$

where N_{spike} denotes the number of emitted spikes, T represents the predefined time window. For remote sensing imagery, the pixel intensity values of the input image are mapped to the corresponding spike counts.

A commonly used Rate coding scheme is Poisson coding [28,29], in which each pixel of the input frame is converted into a Poisson spike train. Since the Poisson distribution characterizes the number of random events occurring within a unit time interval, it aligns well with the notion of spike firing rates. Other rate-based encoding methods include spike density–based encoding [30], population Gaussian Receptive Field (GRF) encoding, and related variants. In contrast, Temporal coding represents information based on precise spike timing, which is more consistent with biological neural mechanisms. Representative temporal encoding approaches include Time-To-First-Spike (TTFS) coding [31], Inter-Spike Interval (ISI) coding [32], Correlation and Synchrony Coding [33], and Temporal Contrast Coding, among others. Among these encoding schemes, Gaussian Receptive Field (GRF) encoding [34] employs a neural population with fixed-parameter Gaussian tuning curves to map continuous input values into spike timing and firing rates, thereby enabling a biologically inspired representation that combines spatial sparsity with temporal encoding. Since the activation of each neuron follows a Gaussian distribution, overlapping sensitivity regions are formed across the neural population. As a result, GRF encoding is particularly effective at capturing multi-scale characteristics of objects in remote sensing imagery, especially for small objects represented by only a few pixels. Moreover, GRF achieves sparse encoding by activating only a limited number of neurons with significant responses, thereby reducing redundant computations. In high-resolution remote sensing scenarios (e.g., large-scale satellite images), this property substantially alleviates the computational burden of network processing while preserving critical feature information.

3 DGRDet

3.1 Network Architecture

The overall architecture of DGRDet is illustrated in Fig. 2. The network consists of five main components, namely: Hierarchical Feedback-based Gaussian Encoding (HFG) module, a backbone network, a spiking feature pyramid network, and classification and regression subnetworks. First, the input images undergo a series of preprocessing operations, including cropping, scaling, and rotation. The preprocessed inputs are then fed into the Gaussian encoding module, where they are converted into spike trains and subsequently passed to the backbone network, SEW-ResNet [35], for target feature extraction. The spiking feature pyramid network takes multiple feature extraction layers generated by SEW-ResNet C_3 (100×100), C_4 (50×50) and C_5 (25×25) as inputs, with one intermediate feature layer C_5 serving as the base feature level. Specifically, an upsampling operation is first applied (indicated by the yellow arrows in Fig. 2), followed by element-wise addition (indicated by the blue blocks) to produce three feature maps at different scales P_3 (100×100), P_4 (50×50) and P_5 (25×25). Subsequently, spiking deconvolution operations (indicated by the blue arrows) are applied to generate two additional feature maps P_6 (13×13) and P_7 (7×7) with smaller spatial resolutions, thereby forming a multi-scale spiking feature representation. Both the classification and regression subnetworks are composed of five sequential convolutional layers with kernel size 3×3. Except for the final layer, the preceding four convolutional layers are each followed by LIF neurons. The classification subnetwork outputs category-wise confidence scores, which are transformed into probability distributions via a sigmoid function. The regression subnetwork extends the conventional four-parameter bounding box offsets by incorporating orientation information, producing a five-dimensional offset vector $[d_x, d_y, d_w, d_h, d_\theta]$. Using this five-parameter representation, rotated bounding boxes are generated and visualized for oriented object detection.

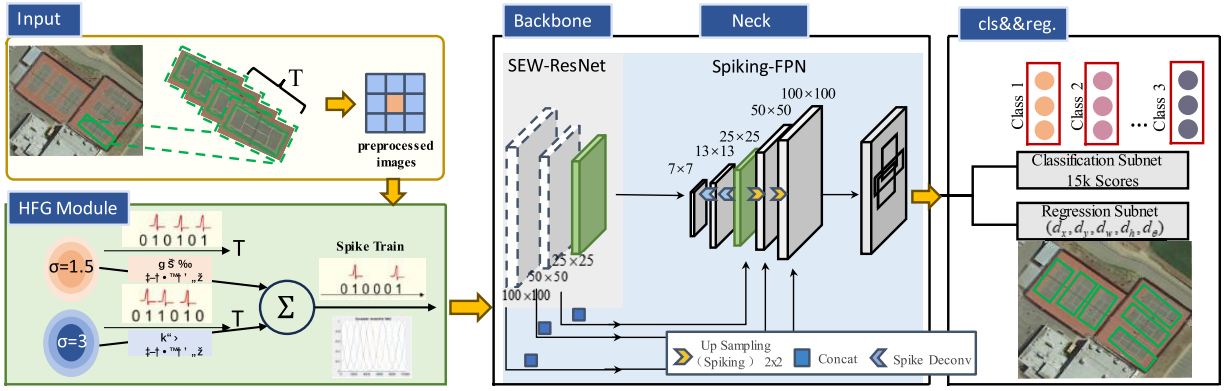

Figure 2: Architecture of the proposed DGRDet detector. In the input image, the object to be detected is tennis court.

3.2 Spiking Neuron

As the fundamental computational unit of neural networks, neurons are responsible for transforming continuous synaptic stimuli into action potential outputs with concrete physical meaning. In ANN, neurons ignore temporal dynamics and propagate information solely in the spatial domain. In contrast, spiking neurons, which incorporate membrane potential dynamics and transmit spatiotemporal information, provide a more biologically plausible modeling paradigm.

However, the ion channel mechanisms of biological neurons are highly complex, as exemplified by models such as the Hodgkin–Huxley model [36] and the Izhikevich model [37]. To effectively simplify computation while remaining as faithful as possible to biological realism, the Leaky Integrate-and-Fire (LIF)

model [38] is widely adopted as a basic computational unit. In this work, DGRDet employs the LIF neuron model proposed by Wu et al. [39], which can be formulated as follows:

$$V_i^{t,f} = \tau V_i^{t-1,f}\left(1 - X_i^{t-1,f}\right) + \sum_j W_{ij}^{f-1} X_j^{t,f-1} \tag{2}$$

$$X_i^{t,f} = H\left(V_i^{t,f} - V_{\text{th}}\right) \tag{3}$$

where the $V_i^{t-1,f}$ is the membrane potential of the i-th spiking neuron in the f layer. As illustrated in Eq. (2), the update of this potential is determined by the decayed accumulation of its preceding state coupled with the integration of the current synaptic input. Specifically, τ denotes the decay factor representing the charge leakage characteristics; when no spike is emitted at the previous timestep (i.e., $X_i^{t-1,f} = 0$), the membrane potential is retained proportionally. The synaptic input component is derived from the weighted summation of the spike sequences $X_j^{t,f-1}$ generated by the neurons in the preceding layer $f-1$. The spike generation process follows the threshold-triggered mechanism defined in Eq. (3), where the Heaviside step function $H(.)$ is employed to evaluate whether the membrane potential reaches the critical threshold V_{th}. Upon meeting the firing condition, the neuron emits a spike and instantaneously resets its membrane potential to a baseline value, thereby completing a full integrate-and-fire cycle.

3.3 Hierarchical Feedback-Based Gaussian Encoding (HFG)

Before introducing the formal equations, we provide an intuitive explanation of HFG. A conventional GRF encoder can be viewed as a static front-end that maps pixel intensity to spike responses with fixed receptive-field locations and widths. Such a design is efficient, but it cannot adjust itself after the detector starts to identify task-relevant structures. HFG turns this one-way process into a closed loop: higher-level spikes indicate which regions and orientations are more informative for the current detection task, and these feedback signals shift, sharpen, or relax the Gaussian receptive fields at the encoder. As a result, the encoder allocates stronger spike responses to object-consistent structures and suppresses background responses that are less useful for oriented localization. In this sense, HFG performs task-driven coarse-to-fine refinement at the input stage rather than relying only on downstream feature extraction to compensate for an initially rigid encoding.

As a static population-based encoding scheme, the conventional Gaussian Receptive Field Encoding (GRF) in SNNs is illustrated in Fig. 3. First, the original input image is normalized as follows:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

where x denotes the input pixel value. Subsequently, each input variable is encoded using n neurons whose receptive fields collectively cover the entire data range, and a width modulation parameter β is introduced to control the spread of the receptive fields. In the receptive field encoding scheme, the center position of each neuron is defined as:

$$c_i = x_{\min} + \frac{2i-3}{2} \cdot \frac{x_{\max} - x_{\min}}{n-2}, \quad i = 1, 2, \ldots, n \tag{5}$$

where, c_i represents the center position of neuron i. This formulation allows the receptive fields of boundary neurons to slightly extend beyond the computational domain, thereby accommodating dynamic variations in pixel values across the image. In the next step, the width of the Gaussian kernel is defined as σ

$$\sigma = \frac{1}{\beta} \cdot \frac{x_{\max} - x_{\min}}{n-2} \tag{6}$$

where, β denotes the width scaling factor of the Gaussian curve. After initializing the parameters of the Gaussian neurons, both the spatial-domain Gaussian responses and the temporal-domain spike encoding can be computed. Specifically, in the spatial domain, given a normalized input $\hat{x}$, the response of the Gaussian neuron i can be defined as:

$$R_i(\hat{x}) = \exp\left(-\frac{(\hat{x} - c_i)^2}{2\sigma^2}\right) \tag{7}$$

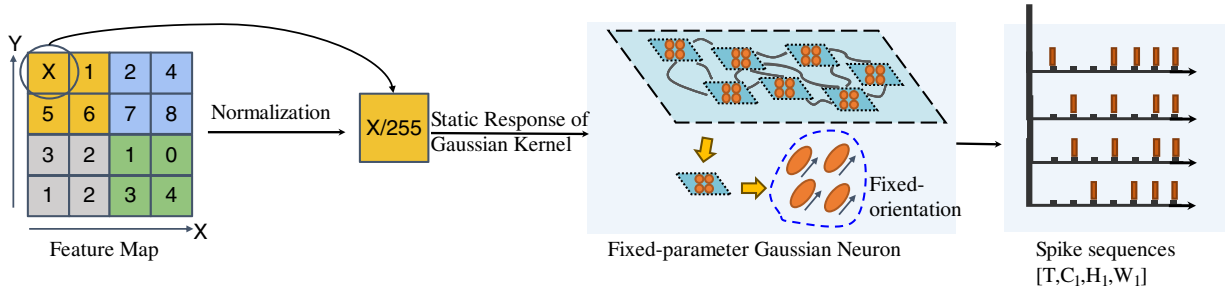

Figure 3: Pulse encoding mechanism based on Gaussian Receptive Fields (GRF). This module maps pixel intensities into initial spike trains via predefined isotropic Gaussian kernels, providing a spatial baseline for subsequent dynamic adjustments.

In the temporal domain, the Gaussian response R_i is mapped to the spike firing time t_i, which is described using a monotonically decreasing function:

$$t_i = T_{\max} \cdot (1 - R_i) \tag{8}$$

However, although Gaussian Receptive Field (GRF) encoding achieves strong robustness and biologically inspired energy efficiency through population coding and sparse spiking, and has been widely adopted in general classification and detection tasks, it exhibits notable limitations when applied to remote sensing object detection. This is primarily because the static and fixed centers and bandwidths of Gaussian kernels prevent the encoded spike sequences from adapting to spatial transformations of the input, such as varying rotation angles of remote sensing objects. Moreover, the receptive field encoding layer is connected to subsequent feature extraction layers only through a feedforward pathway. This unidirectional flow of information prevents the encoding results from being refined by high-level semantic features, thereby limiting the potential improvement in detection accuracy. To address these issues, we draw inspiration from the biological visual cortex, where feedback signals are employed to dynamically adjust the shape and spatial position of neuronal receptive fields in response to retinal stimuli, enabling adaptation to object deformation and rotation. This mechanism aligns well with the intrinsic characteristics of remote sensing image detection. Motivated by this observation, we propose a Hierarchical Feedback-based Gaussian Encoding (HFG) scheme, as illustrated in Fig. 4.

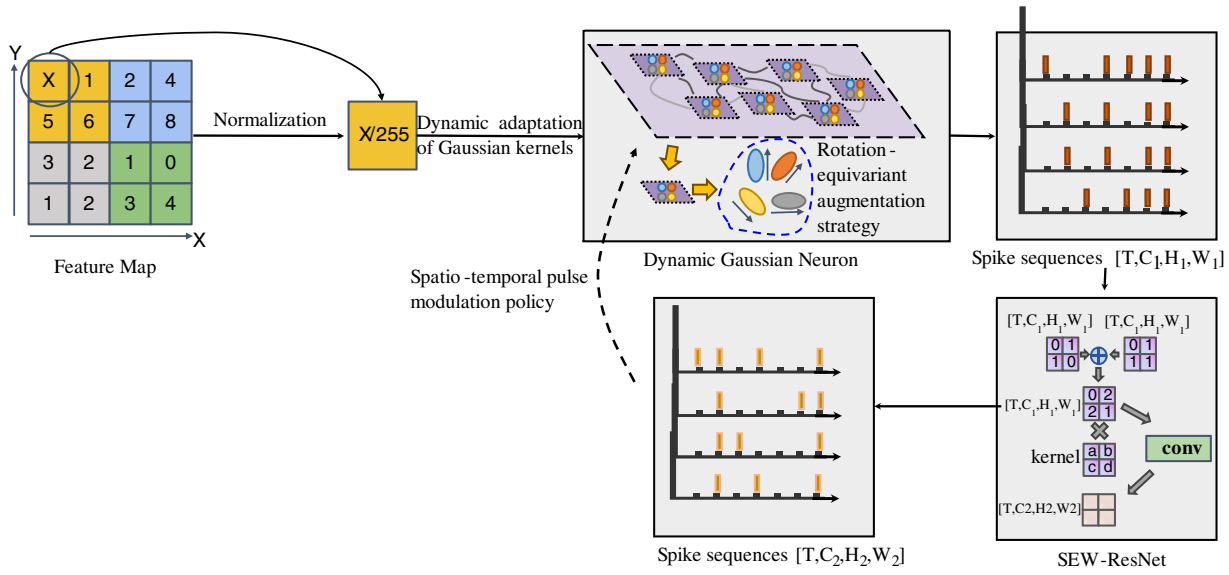

Figure 4: Closed-loop evolution mechanism of Hierarchical Feedback-based Gaussian Encoding (HFG). This mechanism extracts high-level semantic information from the detection network as feedback signals to recursively adjust the spatial parameters of bottom-layer GRFs, achieving task-driven dynamic feature enhancement.

3.3.1 Evolution of Dynamic Receptive Field Adaptation (DRFA)

Based on the conventional Gaussian Receptive Field (GRF) encoding, we introduce two key improvements to enable adaptive feature enhancement driven by hierarchical feedback.

First, we allow both the center and width of the Gaussian kernel to be dynamically adjusted according to feedback signals, which forms the parametric foundation of the proposed adaptive encoding strategy. Specifically, the center of the Gaussian kernel $c_i(t)$ is jointly influenced by the original feedforward input and the top-down feedback signal, formulated as:

$$c_i(t) = \gamma \cdot c_i^s + (1-\gamma) \cdot \sum_j w_{ij}^f \cdot s_j(t-\Delta t) \tag{9}$$

where, c_i^s denotes the initial center position of the Gaussian neuron, w_{ij}^f represents the feedback connection weight, s_j is the spike output of the higher-layer neuron, and γ is a scaling factor controlling the strength of feedback modulation. Through this mechanism, the receptive fields are encouraged to shift toward regions of interest guided by high-level semantic features, enabling task-driven spatial biasing.

Second, since the width of the Gaussian kernel is highly sensitive to object boundaries and deformation characteristics in remote sensing imagery, we further introduce an adaptive optimization strategy for the kernel width:

$$\sigma_i(t) = \sigma_0 \cdot (1 + \upsilon \cdot \| \nabla I(x,y,t) \|) \tag{10}$$

where, σ_0 is the initial kernel width, υ is a normalization factor, and $\nabla I(x,y,t)$ denotes the spatiotemporal gradient magnitude of the input image. At regions with sharp gradient variations (e.g., object edges), υ is reduced to enhance spatial resolution. Conversely, in regions with smoother gradients, υ is increased to suppress background noise. This adaptive modulation enables non-uniform information sampling, effectively balancing detail preservation and noise robustness.

3.3.2 Feedback Spike Modulation (FSM)

After enabling the dynamic adaptation of the Gaussian parameters, we further construct two adaptive enhancement strategies based on the modulated Gaussian representation. The first strategy is a hierarchical feedback mechanism, which is designed to address the modulation of spike-based feedback signals. In our framework, feedback signals from higher network layers are exploited to adjust the Gaussian encoding parameters at lower layers, thereby making the encoding process more compatible with subsequent feature processing. Specifically, when computing the response of a Gaussian neuron, we incorporate not only the feedforward spike activity but also the temporal information of feedback spikes from higher layers for backward modulation. Let the feedback spike time be denoted as t_j^f, the response $R_i^{'}(x,t)$ of the i neuron can then be expressed as:

$$R_i^{'}(x,t) = R_i(x,t) \cdot \Sigma_j \exp\left(-\frac{\left(t - t_j^f\right)^2}{2\tau^2}\right) \tag{11}$$

where, τ denotes the temporal window function. As indicated by Eq. (11), the closer the feedback spike is to the current time t, the stronger the corresponding response enhancement.

3.3.3 Rotation Invariant Enhancement (RIE)

The second strategy focuses on rotation-invariant encoding enhancement for object detection, the enhancement strategy is illustrated in Fig. 5. In biological vision, a well-known phenomenon known as orientation column competition has been extensively studied. The classical Hubel–Wiesel cat visual cortex model [40] revealed that neurons in the primary visual cortex exhibit selectivity to specific edge orientations. These neurons are orderly arranged according to their preferred orientation angles ranging from (0°–180°), forming a characteristic pinwheel-like columnar structure. Moreover, when a particular orientation column is strongly activated (e.g., detecting an edge at 45°), it suppresses the responses of neighboring orientation columns through lateral inhibition. Inspired by this mechanism, we employ clusters of multi-orientation Gaussian kernels to simulate orientation-selective columns, and adopt a maximum-response competition mechanism to model lateral inhibition. In this manner, the biological process of "coverage–suppression–reorganization" in V1 orientation columns is transformed into a computable adaptive module, enabling the proposed model to dynamically optimize orientation representations when encountering rotated objects in remote sensing imagery, analogous to the behavior of biological visual systems.

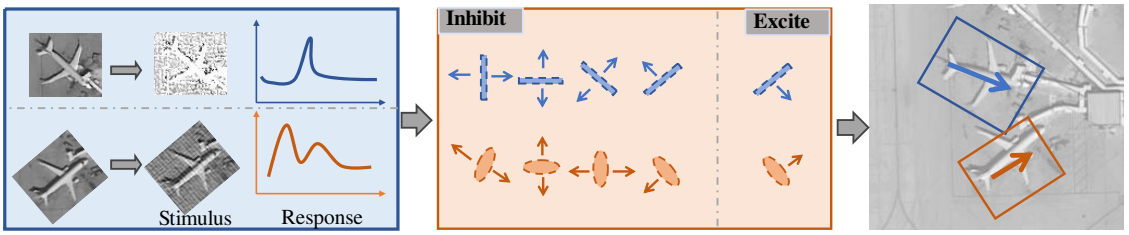

Figure 5: Selective inhibition mechanism of multi-directional Gaussian kernel clusters. This mechanism evaluates the spike intensities across various directional kernels and non-linearly suppresses pseudo-features in non-target orientations, thereby precisely localizing the object's rotation and enhancing the representation of principal directional features.

Specifically, to accommodate the multi-orientation distribution of remote sensing objects, we construct K Gaussian kernels $\left\{R^k\left(x, y, \theta, \right)\right\}$ with predefined orientation angles at each spatial location of the remote sensing feature map, where:

$$\theta_k = \frac{2\pi k}{K}, \quad k = 0, 1, \ldots, K-1 \tag{12}$$

Following the change of θ_k, The response of the k-th orientation-specific kernel can be computed as:

$$R_i^k\left(x, y, t\right) = \exp\left(-\frac{\left(x\cos\theta_k + y\sin\theta_k - c_i\right)^2 + \left(-x\sin\theta_k + y\cos\theta_k\right)^2}{2\sigma_i^2\left(t\right)}\right) \tag{13}$$

According to the biological principle of selective inhibition, only the spike output corresponding to the maximum response is retained at each spatial location, while responses from all other orientations are suppressed. This mechanism enables more accurate inference of object rotation directions during remote sensing object detection. In conventional ANN-based detectors, a similar suppression effect is achieved through Non-Maximum Suppression (NMS), which is also based on eliminating non-maximal responses within local regions. However, NMS operates as a purely post-processing, batch-based engineering heuristic applied to static detection outputs and requires additional design choices for the intersection-over-union (IoU) threshold. In contrast, our approach strictly follows the competitive dynamics of cortical orientation columns, employing real-time spike-based competition and dynamic inhibition. This process is inherently event-driven and more consistent with biologically plausible neural circuit computation, rather than relying on hand-crafted post-processing rules.

3.3.4 Theoretical Analysis of Gaussian Encoding and Receptive Field

To further elucidate the theoretical underpinnings of the proposed Gaussian encoding, it is essential to formalize how the feedback mechanism alters the receptive field geometry and how this connects to oriented object representation. Standard convolutional operations have a fixed, largely axis-aligned receptive field on an isotropic grid. In our framework, spatial features are modeled with a 2D Gaussian distribution characterized by a mean vector μ and a covariance matrix Σ. The feedback mechanism dynamically modulates the receptive field geometry by iteratively updating Σ. By eigendecomposing the covariance matrix, we obtain $\Sigma = R\Lambda R^\top$, where R is an orthogonal rotation matrix whose columns are eigenvectors and Λ is a diagonal matrix of eigenvalues. As Σ is updated through the feedback loop, its decomposition correspondingly changes: Λ controls the spatial extent (scale and aspect ratio) of the receptive field, while R determines its orientation.

Consequently, the receptive field is transformed from a static, axis-aligned region into a dynamic, anisotropic ellipse. This geometric alteration aligns well with the nature of oriented object detection in remote sensing. Targets in satellite imagery (e.g., ships and bridges) often exhibit extreme aspect ratios and arbitrary orientations; an isotropic receptive field may include substantial background clutter, diluting object-specific features. By morphing the receptive field into an oriented Gaussian ellipse via the feedback-updated Σ, feature aggregation is encouraged to align with the object's physical boundaries and rotation angle, increasing the foreground-to-background ratio and benefiting precise oriented bounding box regression.

4 Experiments

We evaluate the performance of the proposed method on the remote sensing object detection task. The spiking neural network is implemented based on the SNN LAVA [41]. In addition to the standard

LIF neuron modeling, spiking residual networks, and analog input processing, we implement the proposed Hierarchical Feedback-based Gaussian Encoding (HFG) module within the LAVA framework. Max-pooling and batch normalization are realized following the implementation strategy described in reference literature [28]. We conduct comprehensive experiments on the large-scale remote sensing datasets DOTA [42] and DIOR [43], including evaluations of detection accuracy, inference speed, and energy consumption. The detailed experimental environment and hardware configurations are summarized in Table 1.

Table 1: Experimental environment.

Category	Environment Configuration
Server	H3C UniServer R4960 G3
CPU	Intel(R) Xeon(R) Gold 5218 @ 2.30 GHz × 64
GPU	NVIDIA Tesla V100-SXM2-16 GB
Memory	253 GiB
Operating System	Debian 11 Bullseye
SNN framework	SNN LAVA
Runtime Environment	Python 3.8

4.1 Datasets

DOTA represents a widely recognized open-source benchmark designed specifically for complex remote sensing imagery. It integrates thousands of pictures collected from an array of sensors and aerial platforms, such as GF-2, JL-1, and Google Earth. The v1.0 release contains 2806 images, featuring significant scale variations from 800 × 800 up to 4000 × 4000 pixels. Following the standard protocol, the images are divided into train, val, and test splits with a proportion of 3:1:2. The dataset captures a rich variety of foreground semantics, containing exactly 188,282 bounding boxes categorized into 15 specific types: Small vehicle (SV), Large vehicle (LV), Plane (PL), Storage tank (ST), Ship (SH), Harbor (HA), Bridge (BR), Ground track field (GTF), Tennis court (TC), Basketball court (BC), Baseball diamond (BD), Soccer ball field (SBF), Roundabout (RA), Swimming pool (SP), and Helicopter (HC).

As a publicly available, large-scale benchmark for optical remote sensing object detection, DIOR comprises 23,463 uniformly sized images (800 × 800 pixels). To faithfully represent real-world and complex remote sensing environments, the images span a wide range of spatial resolutions from 0.5 to 30 m. In total, the dataset provides 192,472 fully annotated, axis-aligned target instances distributed across 20 distinct categories. In this paper, these categories are abbreviated as: airplane (AL), airport (AT), baseball field (BF), basketball court (BC), bridge (BG), chimney (CM), dam (DM), expressway service area (ESA), expressway toll station (ETS), harbor (HB), golf course (GC), ground track field (GTF), overpass (OP), ship (SP), stadium (SD), storage tank (ST), tennis court (TC), train station (TS), vehicle (VH), and windmill (WM).

4.2 Settings

During the direct training of DGRDet, both spike firing dynamics and convolutional computations must be jointly considered. Therefore, a rigorous and well-designed global parameter configuration is critical for overall network convergence and final performance evaluation. In our experiments, the membrane time constant of the LIF neurons is set to =2.0, and the firing threshold is set to 1.0. A sigmoid function is adopted as the surrogate gradient. Given the complexity of remote sensing image features, a decaying strategy is employed for the surrogate gradient width parameter α to avoid training instability. Specifically, α is initially set to 1.2 and gradually decayed to 0.6 as training progresses.

During the training phase, images from the dataset are cropped into 800 × 800 pixel patches along the horizontal and vertical axes, with an overlap stride of 150 pixels between adjacent patches. In the anchor design stage, to accommodate the scale characteristics of remote sensing objects, 21 anchors are predefined at each spatial location on the feature maps (P3, P4, P5, P6, P7). The anchor scales are set to 32, 64, 128, 256, and 512 pixels, respectively. The aspect ratios are set to $\{1/5, 1/3, 1/2, 1, 2, 3, 5\}$, and the scale ratios are set to $\{1, 2^{1/3}, 2^{2/3}\}$.

The model is trained for 183,600 iterations. The initial learning rate is set to 8E−5, linearly increased to 5E−4 over the first 23k iterations and then kept constant. At 65k iterations, the learning rate is decayed to 5E−5.

To avoid ambiguity in the comparison protocol, we clarify the source of reported baseline results. All DGRDet variants and the ablation models are trained and evaluated in our environment. Unless otherwise specified, external baseline results are quoted from the corresponding original papers or official benchmark reports under the same dataset split and evaluation protocol.

4.3 Efficiency of DGRDet

To validate and analyze the efficiency of the proposed method, we investigate the performance of DGRDet in terms of detection efficiency and energy consumption.

4.3.1 Detection Efficiency

We first evaluate detection efficiency, as image processing speed is the most critical criterion for real-time remote sensing applications on terminal devices such as unmanned aerial vehicles. We compare the inference speed of DGRDet with three categories of methods: conventional ANN-based remote sensing object detectors, ANN-to-SNN conversion-based detection algorithms, and directly trained SNN-based detection methods. In addition, to explore the potential of DGRDet for achieving higher detection speed on edge devices, we replace the backbone network with a more lightweight architecture, namely Spiking-MobileNetV2, and measure the corresponding performance metrics. "#Params" denotes the total number of parameters in the model. "Ratio" represents the proportion of the actual runtime of a model relative to the total time required to infer 1000 image patches, where the Ratio metric follows the definition provided by torchstat. Meanwhile, FPS (Frames Per Second) indicates the number of frames processed per second and is measured according to the evaluation protocol described in MMDetection [44].

Efficiency evaluations are conducted on a single GPU with a batch size of 2. The detection efficiency results of the evaluated methods are summarized in Table 2.

The experimental results can be analyzed from the following aspects:

(1) When compared with the conventional ANN-based remote sensing object detection algorithm R3Det, under the same backbone settings (ResNet-50 and SEW-ResNet-50), DGRDet with a time step of $T = 4$ achieves an improvement of 9 fps, corresponding to a relative speed increase of 64.29%. Compared with ANN-to-SNN conversion-based detection methods, DGRDet still exhibits significantly superior detection speed. This advantage mainly stems from the fact that DGRDet operates with only four time steps, whereas conversion-based methods usually require a large number of time steps to preserve detection accuracy. For instance, the compared S3Det algorithm requires 64 time steps to perform remote sensing image detection. In comparison with directly trained SNN-based detection methods, DGRDet achieves a detection speed that is generally comparable. This is because DGRDet incorporates additional encoding and receptive field adjustment mechanisms to ensure detection accuracy, which inevitably introduce a certain amount of

computational overhead during inference. Nevertheless, the overall speed cost remains acceptable, and the method maintains a favorable balance between efficiency and accuracy.

(2) When replacing the backbone network with a lightweight architecture, such as the spiking version of MobileNetV3, the detection speed can be further increased to 40 fps, while the actual runtime ratio is reduced to only 21.14%. Meanwhile, the total number of network parameters is limited to 168 MB. These results demonstrate that DGRDet is promising for deployment on embedded platforms such as unmanned aerial vehicles and mobile robots, indicating strong potential for practical and industrial applications.

Table 2: Speed comparison on DOTA.

Model	Backbone	Image Size	Spike	DOTA		
				#Params	Ratio	Speed
ANN						
R3Det [23]	ResNet50	800 * 800	×	485 MiB	88.52%	14 fps
SCRDet [45]	ResNet50	800 * 800	×	452 MiB	71.20%	10 fps
R2CNN [46]	ResNet50	600 * 600	×	353 MiB	93.60%	–
RRPN [47]	ResNet50	600 * 600	×	348 MiB	94.30%	5 fps
RetinaNet-R [48]	ResNet50	800 * 800	×	378 MiB	82.85%	12 fps
MobileDet-R [42]	MobileNetV3	300 * 300	×	96 MiB	31.78%	41.5 fps
GiraffeDet-R [49]	S2D Chain	300 * 300	×	137 MiB	35.32%	35 fps
OR-CNN [50]	ResNet50	800 * 800	×	368 MiB	37.31%	15.3 fps
DFDet [26]	ResNet50	800 * 800	×	392 MiB	36.82%	23.4 fps
YOLOv10-L [51]	CSPNet	300 * 300	×	152 MiB	40.73%	32 fps
ANN to SNN						
S3Det (T = 64) [52]	ResNet50	800 * 800	√	462 MiB	64.47%	20 fps
Spiking-RetinaNet-R (T = 64)	ResNet50	800 * 800	√	415 MiB	54.28%	24 fps
Directly Trained SNN						
Meta-spikerformer-R (T = 4) [13]	Conv+ViT	800 * 800	√	475 MiB	83.28%	24 fps
Spiking-YOLO-R (T = 4) [6]	DarkNet	300 * 300	√	172 MiB	33.16%	45 fps
DGRDet (T = 4)	SEW-ResNet18	800 * 800	√	362 MiB	35.47%	27 fps
	SEW-ResNet34	800 * 800	√	385 MiB	37.57%	24 fps
	SEW-ResNet50	800 * 800	√	487 MiB	42.29%	23 fps
	S-MobileNetV3	300 * 300	√	168 MiB	21.14%	40 fps

4.3.2 Energy Consumption Ablation Study

The energy efficiency advantage of SNN mainly arises from their event-driven spatiotemporal dynamics, where accumulation calculations (AC) and data transmissions are triggered only when spikes are emitted. For the proposed method, the biologically inspired hierarchical recursive Gaussian encoding adopts a population coding strategy. Although multiple neurons may participate in spiking activity within a single time step, the membrane potential updates can be completed with an extremely small number of time steps, thereby effectively controlling the overall energy consumption. To better demonstrate the significant energy efficiency of DGRDet, we evaluate the energy consumption of three categories of methods using SNN

Toolbox: conventional ANN-based remote sensing object detection algorithms, ANN-to-SNN conversion-based detection methods, and the proposed DGRDet. During energy estimation, each time step is defined as 1 ms, following the 1 kHz synchronous signal assumption adopted by Merolla et al. [53]. According to the analysis by Horowitz [54], the energy cost of a single Float32 multiply–accumulate (MAC) operation is 4.6 pJ, while that of an accumulate (AC) operation is 0.9 pJ. Considering that DGRDet accepts analog-valued image inputs, the first network layer is modeled using MAC operations, while all subsequent layers are implemented with AC operations. FLOPs refers to floating-point operations (multiply-accumulate) used in ANNs. SOPs refers to synaptic operations (accumulations/additions) driven by spikes in SNNs. In addition, an ablation study is conducted to compare the proposed Hierarchical Feedback-based Gaussian Encoding (HFG) with the conventional Gaussian Receptive Field (GRF) encoding, in order to further verify the effectiveness of the internal encoding mechanism in terms of energy efficiency.

To ensure fair and consistent comparisons, all models are trained and evaluated under strictly identical experimental settings. The energy consumption results of different detection algorithms, as well as the ablation analysis within DGRDet, are summarized in Table 3.

Table 3: Energy consumption comparison on DOTA.

	Data Type	Input	FLOPs	SOPs	Spiking Rate	Energy (J)	Power (W)
R3Det	Float 32bit	800 × 800	4.334E+11	–	–	1.994	178
S3Det (T = 64)	Float 32bit	800 × 800	–	4.275E+11	24.32%	9.36E−02	1.46
DGRDet (GRF) T = 4	Float 32bit	800 × 800	–	1.126E+10	23.81%	2.41E−03	0.60
DGRDet (HFG) T = 4	Float 32bit	800 × 800	–	1.175E+10	18.69%	1.98E−03	0.49

According to the energy consumption results, DGRDet equipped with the conventional Gaussian Receptive Field (GRF) encoding consumes 2.41×10^{-3} J when using 32-bit floating-point input precision, which corresponds to only 0.12% of the energy consumed by the ANN-based remote sensing detector R3Det with an identical network structure. The corresponding power consumption is 0.60 W, accounting for merely 0.82% of that of R3Det. These results clearly demonstrate the significant low-power advantage of SNN-based models. Even when compared with the ANN-to-SNN conversion-based method S3Det, DGRDet still exhibits a pronounced advantage in terms of power efficiency.

Furthermore, in the internal ablation study, all other variables are kept identical. When adopting the proposed Hierarchical Feedback-based Gaussian Encoding (HFG), the overall spike firing rate of the model is reduced to 18.69%, representing an approximately 5.12% decrease compared with the GRF-based counterpart. This reduction is primarily attributed to the enhanced rotation-invariant encoding mechanism, which effectively suppresses erroneous spike activations induced by incorrect object orientations, thereby substantially alleviating redundant spike firing. In addition, the model equipped with HFG achieves the lowest energy consumption and power dissipation among all compared configurations, further validating the overall effectiveness and efficiency of the proposed encoding strategy.

It should be noted that, due to the limited availability of deployable neuromorphic hardware in our current experimental environment, we did not perform direct on-chip energy measurements in this study. Instead, the energy values reported in this subsection are theoretical estimates computed under a unified operation-based model from MAC/AC/SOP counts and spike statistics, which is a common practice in the SNN literature. Therefore, these results should be interpreted as analytical estimates for relative comparison

rather than board-level measurements on a specific neuromorphic processor. Actual hardware energy may differ because of memory access, routing, synchronization, I/O, and device-specific mapping overheads.

4.4 High-Precision Detection Experiments

The pronounced performance gap in detection accuracy has long been one of the most critical challenges hindering the practical adoption of Spiking Neural Networks (SNNs). Although several recent studies have demonstrated that SNNs can achieve lossless or even superior performance compared to ANNs on natural image benchmarks [6,13], their effectiveness on remote sensing datasets remains limited. As an exploratory study of directly trained deep SNNs for remote sensing object detection, we conduct comprehensive accuracy comparison experiments to evaluate the performance of DGRDet. Precision experiments are conducted using four GPUs for both training and inference, with a batch size of 8. For a fair comparison with classical ANN-based object detectors, all compared models adopt ResNet-50 as the backbone network. We evaluate DGRDet on both the DOTA and DIOR datasets under two distinct temporal configurations, specifically setting the number of time steps to 1 and 4. We compare the detection accuracy of DGRDet against various commonly used baseline methods. The quantitative results on the DOTA dataset are summarized in Table 4, while the detection accuracy on the DIOR dataset is illustrated in Fig. 6.

Table 4: Evaluation of the oriented bounding box (OBB) task on the DOTA testing set. The bolded numbers indicate the highest accuracy rate under the current category. The abbreviations at the first line can be referred to the introduction of DOTA in Section 4.1.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HB	SP	HC	mAP
One-stage																
RetinaNet-R	88.92	67.67	33.55	56.83	66.11	73.28	75.24	90.87	73.95	75.07	43.77	56.72	51.05	55.86	21.46	62.02
DAL [55]	88.68	76.55	45.08	66.80	67.00	**76.76**	**79.74**	90.84	79.54	78.45	57.71	62.27	69.05	73.14	60.11	71.44
R^3Det-50	89.30	80.29	46.21	65.07	70.51	73.38	77.42	90.83	80.59	82.26	59.29	58.25	57.75	65.90	55.31	70.16
Two-stage																
SCRDet	**89.98**	80.65	**52.09**	68.36	64.52	60.32	72.41	90.85	**87.94**	**86.86**	65.02	66.68	66.25	68.24	65.21	72.36
R2CNN	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	59.69	52.84	53.08	51.94	53.58	61.01
ICN [56]	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.20	53.64	62.90	67.02	64.17	50.23	68.16
CAD-Net [57]	87.80	**82.40**	49.40	**73.50**	71.10	63.50	76.60	90.90	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
ANN2SNN																
S3Det-50 (T = 512)	89.83	69.45	36.05	66.50	64.34	75.10	68.11	**92.56**	77.33	80.28	57.55	68.88	64.88	66.11	54.84	68.79
Directly Trained SNN																
DGRDet (T = 1)	88.67	80.44	42.35	58.72	71.03	63.89	65.20	86.45	77.10	77.88	58.55	60.12	73.25	55.60	61.90	67.41
DGRDet (T = 4)	86.26	81.30	43.21	62.85	**76.40**	68.12	69.33	90.05	80.58	81.20	62.45	63.10	**74.66**	56.89	68.55	**70.33**

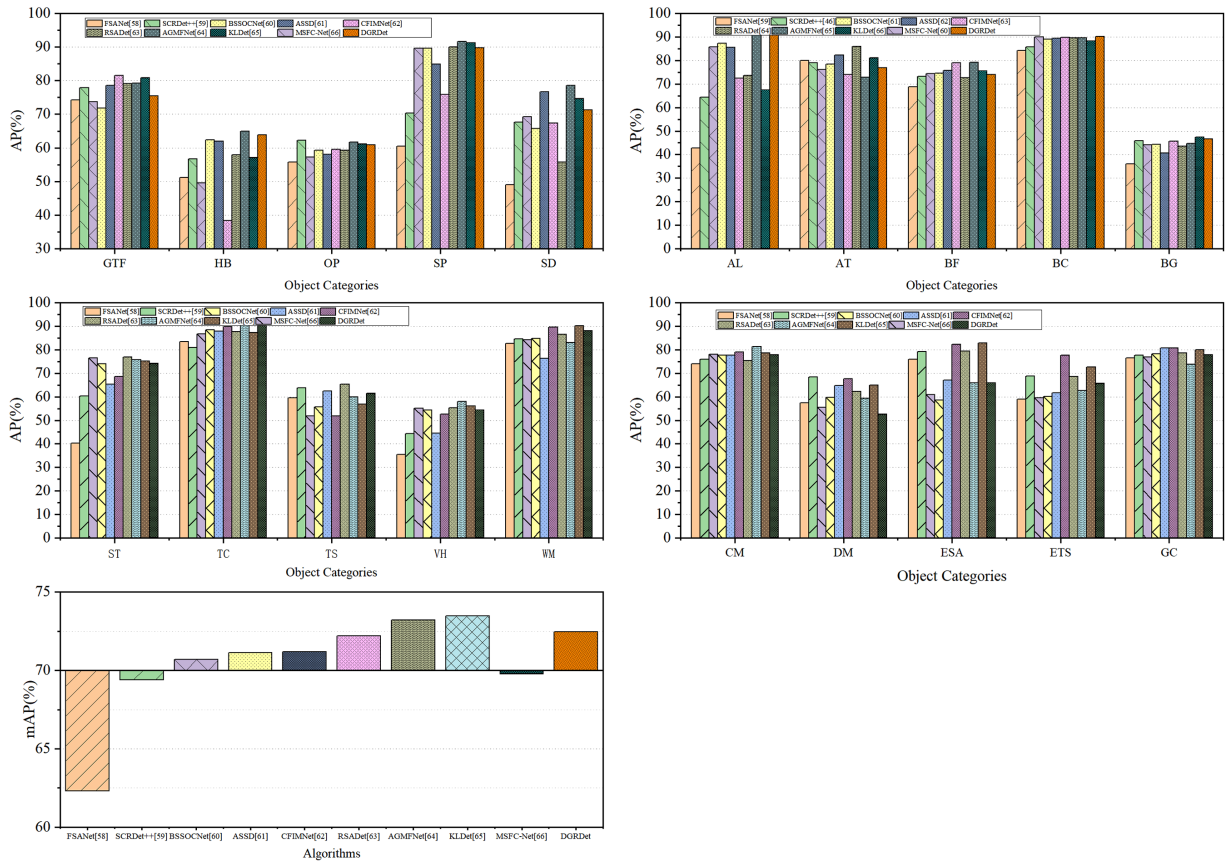

Figure 6: Evaluation of the HBB task on the DIOR testing set.

The quantitative results demonstrate that DGRDet achieves competitive performance on the DOTA dataset. Specifically, with only 4 time steps, DGRDet attains a mAP of 70.33%, outperforming the ANN-to-SNN conversion-based method S3Det, which requires 512 time steps, by nearly 2% mAP. This highlights the effectiveness of the proposed direct-training SNN paradigm in preserving detection accuracy under extremely short temporal dynamics. Notably, the detection performance of DGRDet surpasses a large number of classical ANN-based detectors and is only 2.03% mAP lower than the state-of-the-art ANN detector SCRDet. Except for scenarios that demand extremely high accuracy, the detection performance achieved by S3Det is already sufficient for most practical deployment scenarios on edge and terminal devices. Therefore, we argue that the slight accuracy gap compared to the SOTA ANN detector is acceptable given the substantial gains in efficiency and energy consumption. Furthermore, DGRDet achieves state-of-the-art results on several challenging object categories, including Small Vehicle (SV) and Harbor (HB). These categories are characterized by large aspect ratios and highly diverse orientations, which validates the proposed method's strong capability in handling multi-rotation remote sensing objects. The results also indicate that DGRDet effectively captures high-level semantic contour features under complex geometric transformations. Qualitative visualization results on the DOTA dataset are presented in Fig. 7.

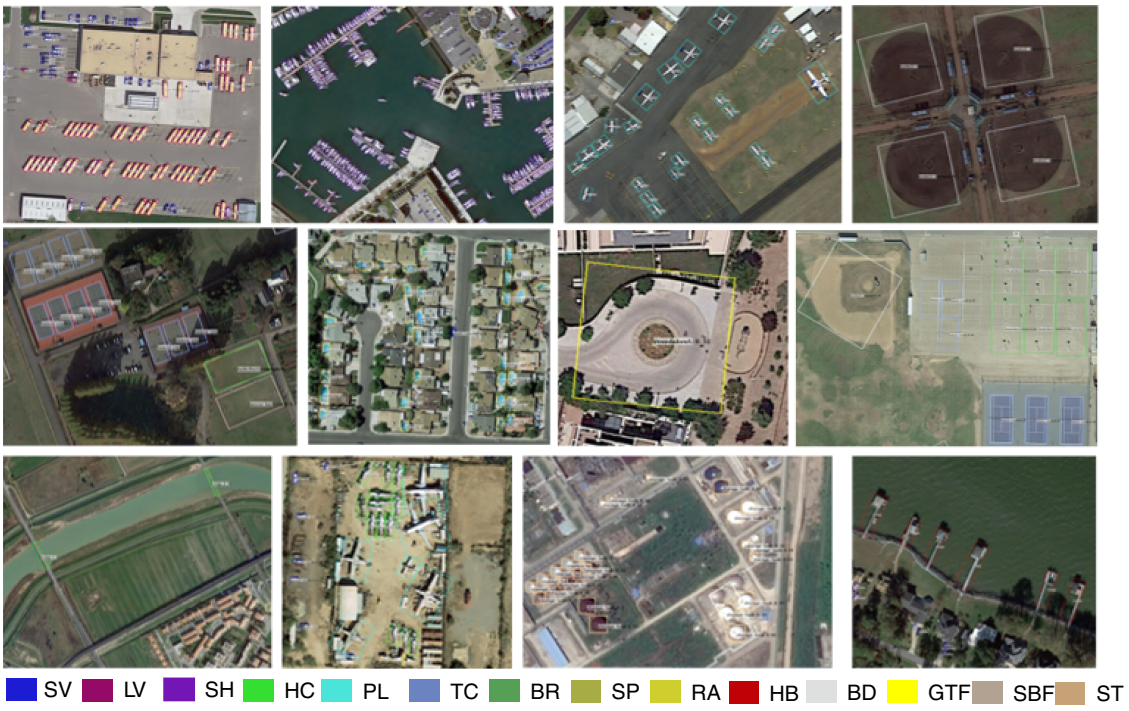

Figure 7: Visualization of DGRDet on DOTA. The short names for categories are defined as: SV, Small vehicle; LV, Large vehicle; SH, Ship; HC, Helicopter; PL, Plane; TC, Tennis court; BR, Bridge; SP, Swimming pool; RA, Roundabout; HB, Harbor; BD, Baseball diamond; GTF, Ground field track; SBF, Soccer-ball field; ST, Storage tank; and BC, Basketball court.

Evaluated on the DIOR dataset with horizontal bounding box (HBB) annotations, our proposed DGRDet achieves an mAP of 72.46%, ranking third in accuracy among the 10 compared models, the comparative models include FSANet [58], SCRDet++ [59], BSSOCNet [60], ASSD [61], CFIMNet [62], RSADet [63], AGMFNet [64], KLDet [65], and MSFC-Net [66]. A detailed category-wise analysis reveals that DGRDet attains the highest precision among all baselines on targets with distinct contour features, such as airplane (91.7%), basketball court (90.2%), and tennis court (90.2%). However, it exhibits performance shortcomings on categories characterized by complex morphologies or severe background interference, such as dam (52.7%) and bridge (46.6%). Admittedly, the discrete nature of spiking signals and the inherent optimization challenges associated with surrogate gradient computation make it fundamentally difficult to further elevate the accuracy of DGRDet. Against this backdrop, DGRDet successfully overcomes these training bottlenecks while preserving the inherent low-power advantage of SNNs, achieving an overall detection accuracy exceeding 72%. This overall performance is comparable to, and in some aspects even surpasses, mainstream high-precision Artificial Neural Network (ANN) models. This objectively demonstrates the effectiveness and technical feasibility of our proposed algorithmic mechanism in handling complex remote sensing object detection tasks.

Analysis of the Remaining Gap to ANN. Although DGRDet reaches a competitive overall mAP on DOTA, the remaining gap to ANN detectors is not uniform across categories. Relative to SCRDet, the largest deficits of DGRDet (T = 4) appear on swimming pool (−11.35 AP), bridge (−8.88), basketball court (−7.36), storage tank (−5.66), and ground track field (−5.51), whereas gains are observed on small vehicle (+11.88), harbor (+8.41), large vehicle (+7.80), and helicopter (+3.34). This pattern suggests that the proposed encoding is particularly helpful for categories with strong directional cues or large orientation variation, but it is less

effective for categories that rely more on fine interior texture, dense local details, or richer long-range context. We therefore attribute the remaining gap mainly to three factors: (1) information discretization under low-timestep spiking dynamics still weakens fine-grained appearance modeling; (2) the current convolutional SNN backbone does not model global context as strongly as specialized ANN detectors with richer attention or denoising modules; and (3) HFG improves the input representation, but it does not yet remove all downstream bottlenecks in feature aggregation and box regression. This interpretation is also consistent with the weaker DIOR performance on categories such as bridge and dam, where background interference and structural ambiguity remain challenging.

4.5 Ablation Study on Simulation Time Step

In Spiking Neural Networks, the simulation time step T is a crucial hyperparameter, as it dictates the size of the temporal window for integrating spatiotemporal information. To justify the selection of T = 4 as the optimal configuration for DGRDet, we conduct a comprehensive ablation study across various time steps ($T \in \{1, 2, 4, 8, 16\}$). We evaluate the mAP on the DOTA dataset and quantify the corresponding computational costs, encompassing Synaptic Operations (SOPs), theoretical energy consumption, and inference speed (FPS). As the time step increases geometrically, the required GPU memory grows rapidly. Constrained by our hardware environment, we configure the experiments using 8 GPUs with a batch size of 16. Under this setup, when T = 16, the training process suffers from Out-of-Memory (OOM) issues on standard GPUs due to the heavy memory footprint of BPTT, further proving the impracticality of overly large time steps. The detailed experimental results are summarized in Table 5, and the performance-efficiency trade-off is visually illustrated in Fig. 8.

Table 5: Performance, computational cost, and speed of DGRDet under various time steps.

Time Step (T)	mAP_{50} (%)	Speed (fps)	SOPs	Spiking Rate	Energy (J)	Power (W)
1	62.28	65	2.827E+09	22.24%	6.28E−04	0.63
2	67.17	42	5.633E+09	23.28%	1.18E−03	0.60
4	70.33	23	1.126E+10	23.81%	2.41E−03	0.60
8	70.75	12	2.252E+10	22.69%	5.11E−03	0.63
16	70.85	6	4.504E+10	24.57%	1.11E−02	0.69

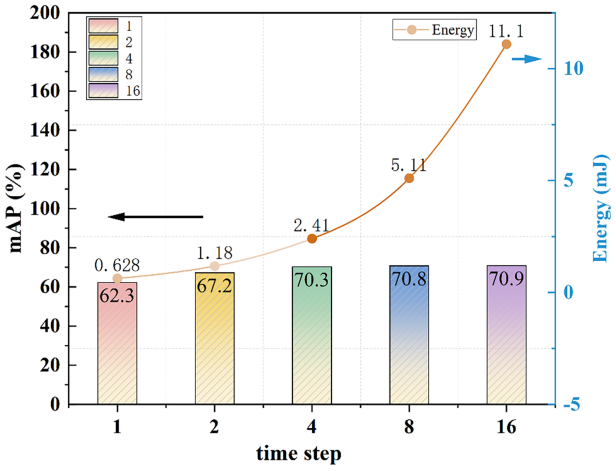

Figure 8: Detection accuracy and energy consumption of DGRDet under different time-steps.

As shown in Table 5, when the time step increases from $T = 1$ to $T = 4$, the mAP50 jumps from 62.28% to 70.33%. This significant performance improvement indicates that sufficient time steps are essential for SNNs to accurately locate targets in complex remote sensing images.

However, obvious accuracy saturation occurs when $T > 4$. Increasing T from 4 to 8 brings only a minor mAP gain of 0.42%, and accuracy almost stops growing at $T = 16$. In sharp contrast, the computational cost grows almost linearly with T. For example, compared to $T = 4$, setting $T = 8$ almost doubles the energy consumption (5.11E−03 vs. 2.41E−03 J) and halves the inference speed. Additionally, since backpropagation unrolls the network along the time axis, setting $T = 16$ causes a sharp increase in memory usage, greatly increasing training difficulty.

Currently, remote sensing detection algorithms usually need to be deployed on edge platforms like UAVs or satellites, which have strict limits on power and real-time processing. As shown in Fig. 8, $T = 4$ achieves near-peak accuracy while maintaining low energy consumption and acceptable latency, achieving the best trade-off between accuracy and efficiency. Any larger T brings significant energy increases, failing to meet the actual efficiency needs of remote sensing edge applications. Therefore, $T = 4$ is the most reasonable configuration for DGRDet.

4.6 Ablation Study on the HFG Encoding Module

To isolate the contribution of the proposed encoding module from the rest of the detector, we add a dedicated component-wise ablation study in which the backbone, spiking FPN, detection heads, optimizer, training schedule, input resolution, and simulation time step are all kept fixed, and only the encoder is varied. Specifically, we compare a static GRF encoder, partial HFG variants that enable only one component at a time, and the full HFG. Here, the HFG encoder is decomposed into three functional components: Dynamic Receptive Field Adaptation (DRFA), Feedback Spike Modulation (FSM), and Rotation Invariant Enhancement (RIE). This protocol makes it possible to determine whether the observed gains arise from the task-adaptive encoding design itself rather than from changes in the downstream detection architecture.

As shown in Table 6, all three components of the proposed HFG encoder contribute positively to the final detection performance. Starting from the static GRF baseline, enabling DRFA alone improves the mAP from 68.24% to 68.97%, indicating that adaptive receptive field adjustment can better match the scale and spatial structure variations of remote sensing objects. Enabling FSM alone further increases the mAP to 69.18%, suggesting that top-down spike modulation improves the compatibility between the encoding process and downstream hierarchical feature extraction.

Table 6: Component-wise ablation of the proposed HFG encoder on DOTA (all downstream detector components are fixed).

Variant	DRFA	FSM	RIE	mAP (%)	Spike Rate (%)	Energy (J)	Speed (FPS)
GRF baseline	×	×	×	68.24	23.81	2.41E−03	26
GRF+DRFA	√	×	×	68.97	22.94	2.30E−03	24
GRF+FSM	×	√	×	69.18	22.16	2.22E−03	24
GRF+RIE	×	×	√	69.56	20.87	2.09E−03	24
Full HFG	√	√	√	70.33	18.69	1.98E−03	23

Among the three individual components, RIE yields the largest single-module gain, improving the mAP to 69.56% while also reducing the spike rate to 20.87% and the analytical energy consumption to 2.09E−03 J. This trend is consistent with the characteristics of the DOTA dataset, where many targets exhibit

arbitrary orientations and elongated structures. By suppressing non-target directional responses, RIE not only improves orientation-aware representation but also effectively reduces redundant spike firing.

When the three components are integrated, the full HFG achieves the best overall result, reaching 70.33% mAP, with the lowest spike rate (18.69%) and lowest energy consumption (1.98E−03 J) among all compared variants, while maintaining a comparable inference speed (23 FPS). These results indicate that DRFA, FSM, and RIE are complementary rather than redundant, and the observed gain mainly comes from the coordinated adaptive encoding design itself.

4.7 Discussion

Discussion on Generalization. Although the current experiments are conducted on DOTA and DIOR, the proposed HFG mechanism is not specifically designed for a particular dataset, category taxonomy, or detection head. Instead, HFG operates at the encoding stage, where receptive-field position, scale, and orientation are adaptively adjusted before backbone feature extraction. This makes the mechanism fundamentally feature-oriented rather than dataset-specific. Since many remote sensing detection benchmarks share common challenges such as multi-scale targets, arbitrary object orientations, cluttered backgrounds, and large appearance variations, the proposed encoding strategy is expected to be transferable to other remote sensing detection scenarios as well. Nevertheless, we acknowledge that this claim is currently supported by the mechanism design and the results on two representative benchmarks, rather than by exhaustive cross-dataset experiments. A broader validation on additional remote sensing datasets will therefore be an important part of our future work.

Practicality for UAV and Edge Deployment. While this study primarily validates DGRDet through simulation environments, its design inherently aligns with the strict Size, Weight, and Power (SWaP) constraints of real-world UAVs and edge devices. The practicality of deploying DGRDet stems from its extreme temporal sparsity. Traditional artificial neural networks (ANNs) rely on dense Multiply-Accumulate (MAC) operations, causing severe thermal throttling and battery drainage on micro-drones. In contrast, DGRDet employs event-driven accumulation (AC) operations. Specifically, as demonstrated in our ablation studies, the proposed HFG encoder effectively suppresses background noise and non-target responses, driving the network's overall spike firing rate down to approximately 18.69% and analytical energy consumption to 1.98E−03 J per inference.

In practical hardware deployment, this high degree of sparsity directly translates to minimized memory access frequency and reduced dynamic power consumption. Furthermore, the functional modules of HFG (e.g., dynamic receptive fields and feedback modulation) are implemented through localized membrane potential dynamics rather than complex global attention matrices, ensuring structural compatibility with emerging neuromorphic chips (such as Intel Loihi 2) and edge-accelerated FPGAs. Future work will focus on the hardware-in-the-loop quantization and physical deployment of DGRDet to measure real-world latency and milliwatt-level power efficiency on airborne platforms.

5 Conclusions

In this work, we investigate the feasibility of applying a directly trained deep Spiking Neural Network (SNN) to remote sensing image object detection. Considering that remote sensing object detection requires the extraction of complex, rotation-aware, and multi-scale features, we introduce a novel spiking biologically inspired hierarchical feedback-based Gaussian encoding mechanism. Specifically, we first theoretically analyze the limitations of conventional Gaussian encoding schemes and design a biologically hierarchical feedback-based encoding strategy inspired by principles of biological vision. Furthermore, to address the difficulty of detecting remote sensing objects with arbitrary orientations, we propose a rotation-invariance

enhancement strategy at the encoding stage, which effectively reduces redundant spike firing during the detection process. The directly trained deep DGRDet framework facilitates straightforward deployment on lightweight and resource-constrained devices. Experimental results demonstrate that, with an extremely small number of time steps, the proposed model achieves performance comparable to ANN-based counterparts with the same network structure, while consuming significantly less power than both ANN models and ANN-to-SNN conversion methods. Moreover, DGRDet exhibits superior performance on authoritative remote sensing benchmarks.

Nevertheless, we have not yet optimized the subsequent network architecture in this study. In future work, we plan to further explore the topological designs of core feature extraction networks in SNNs and investigate their integration with neuromorphic hardware, thereby enabling rapid and efficient real-world deployment.

Acknowledgement: The authors would like to express their gratitude to Wei Guo for providing valuable feedback, and to Yun Huang for their support.

Funding Statement: This research was funded by the National Key R&D Program of China Grant No. 2022YFB4500900.

Author Contributions: Li Chen was responsible for the conception and planning of the experimental ideas, and led the design of the experiments. He also participated in data analysis and interpretation of results. Fan Zhang undertook the task of const ructing the theoretical model and conducted a com-prehensive review of the relevant literature. Guangwei Xie focused on the data collection process and the execution of the experiments. Yanzhao Gao was in charge of organizing the theoretical knowledge into mathematical formulations. Xiaofeng Qi took the lead in drafting the initial manuscript and subsequent revisions, ensuring that the paper's structure was logical and the language was fluent. Mingqian Sun managed the collection and organization of preliminary research materials. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support this study are available from authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nikouei M, Baroutian B, Nabavi S, Taraghi F, Aghaei A, Sajedi A, et al. Small object detection: a comprehensive survey on challenges, techniques and real-world applications. Intell Syst Appl. 2025;27(1):200561. doi:10.1016/j.iswa.2025.200561.
2. Maass W. Networks of spiking neurons: the third generation of neural network models. Neural Netw. 1997;10(9):1659–71. doi:10.1016/S0893-6080(97)00011-7.
3. Shrestha SB, Orchard G. SLAYER: spike layer error reassignment in time. arXiv:1810.08646. 2018.
4. Cao Y, Chen Y, Khosla D. Spiking deep convolutional neural networks for energy-efficient object recognition. Int J Comput Vis. 2015;113(1):54–66. doi:10.1007/s11263-014-0788-3.
5. Hunsberger E, Eliasmith C. Training spiking deep networks for neuromorphic hardware. arXiv:1611.05141. 2016.
6. Kim S, Park S, Na B, Yoon S. Spiking-YOLO: spiking neural network for energy-efficient object detection. Proc AAAI Conf Artif Intell. 2020;34(7):11270–7. doi:10.1609/aaai.v34i07.6787.
7. Li Y, He X, Dong Y, Kong Q, Zeng Y. Spike calibration: fast and accurate conversion of spiking neural network for object detection and segmentation. arXiv:2207.02702. 2022.
8. Jin X, Zhang M, Yan R, Pan G, Ma D. R-SNN: region-based spiking neural network for object detection. IEEE Trans Cogn Dev Syst. 2024;16(3):810–7. doi:10.1109/tcds.2023.3311634.

9. Everingham M, Zisserman A, Williams CKI. The 2005 PASCAL visual object classes challenge. In: Machine learning challenges workshop. Berlin/Heidelberg, Germany: Springer; 2005. p. 117–76. doi:10.1007/11736790_8.
10. Qu J, Gao Z, Zhang T, Lu Y, Tang H, Qiao H. Spiking neural network for ultra-low-latency and high-accurate object detection. arXiv:2306.12010. 2023.
11. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer vision—ECCV 2014. Cham, Switzerland: Springer International Publishing; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.
12. Su Q, Chou Y, Hu Y, Li J, Mei S, Zhang Z, et al. Deep directly-trained spiking neural networks for object detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 6532–42. doi:10.1109/ICCV51070.2023.00603.
13. Yao M, Hu J, Hu T, Xu Y, Zhou Z, Tian Y, et al. Spike-driven transformer V2: meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. arXiv:2404.03663. 2024.
14. Xia GS, Bai X, Ding J, Zhu Z, Belongie S, Luo J, et al. DOTA: a large-scale dataset for object detection in aerial images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 3974–83. doi:10.1109/cvpr.2018.00418.
15. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020.
16. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44. doi:10.1109/CVPR.2017.106.
17. Girshick R. Faster R-CNN. In: Proceedings of the IEEE International Conference On Computer Vision; 2015 Dec 11–18; Santiago, Chile. p. 1440–448.
18. Ding J, Xue N, Long Y, Xia GS, Lu Q. Learning RoI transformer for detecting oriented objects in aerial images. arXiv:1812.00155. 2018.
19. Xie X, Cheng G, Wang J, Yao X, Han J. Oriented R-CNN for object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 3500–9. doi:10.1109/iccv48922.2021.00350.
20. Pu Y, Wang Y, Xia Z, Han Y, Wang Y, Gan W, et al. Adaptive rotated convolution for rotated object detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 6566–77. doi:10.1109/ICCV51070.2023.00606.
21. Yu H, Tian Y, Ye Q, Liu Y. Spatial transform decoupling for oriented object detection. Proc AAAI Conf Artif Intell. 2024;38(7):6782–90. doi:10.1609/aaai.v38i7.28502.
22. Yao Y, Cheng G, Wang G, Li S, Zhou P, Xie X, et al. On improving bounding box representations for oriented object detection. IEEE Trans Geosci Remote Sensing. 2023;61:1–11. doi:10.1109/tgrs.2022.3231340.
23. Yang X, Yan J, Feng Z, He T. R3Det: refined single-stage detector with feature refinement for rotating object. Proc AAAI Conf Artif Intell. 2021;35(4):3163–71. doi:10.1609/aaai.v35i4.16426.
24. Sun P, Zheng Y, Wu W, Xu W, Bai S, Lu X. Learning critical features for arbitrary-oriented object detection in remote-sensing optical images. IEEE Trans Instrum Meas. 2024;73:5015112. doi:10.1109/TIM.2024.3378265.
25. Li W, Chen Y, Hu K, Zhu J. Oriented RepPoints for aerial object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 1819–28. doi:10.1109/CVPR52688.2022.00187.
26. Xie X, Cheng G, Rao C, Lang C, Han J. Oriented object detection via contextual dependence mining and penalty-incentive allocation. IEEE Trans Geosci Remote Sens. 2024;62:5618010. doi:10.1109/TGRS.2024.3385985.
27. Zeng Y, Chen Y, Yang X, Li Q, Yan J. ARS-DETR: aspect ratio-sensitive detection transformer for aerial oriented object detection. IEEE Trans Geosci Remote Sens. 2024;62:5610315. doi:10.1109/TGRS.2024.3364713.
28. Rueckauer B, Lungu IA, Hu Y, Pfeiffer M, Liu SC. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. Front Neurosci. 2017;11:682. doi:10.3389/fnins.2017.00682.

29. Diehl PU, Neil D, Binas J, Cook M, Liu SC, Pfeiffer M. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In: 2015 International Joint Conference on Neural Networks (IJCNN); 2015 Jul 12–17; Killarney, Ireland. p. 1–8. doi:10.1109/IJCNN.2015.7280696.
30. Gerstner W, Kistler WM. Spiking neuron models. Cambridge, UK: Cambridge University Press; 2002.
31. Gollisch T, Meister M. Rapid neural coding in the retina with relative spike latencies. Science. 2008;319(5866):1108–11. doi:10.1126/science.1149639.
32. Park S, Kim S, Choe H, Yoon S. Fast and efficient information transmission with burst spikes in deep spiking neural networks. In: 2019 56th ACM/IEEE Design Automation Conference (DAC); 2019 Jun 2–6; Las Vegas, NV, USA. p. 2019:1–6.
33. Gerstner W, Kistler WM, Naud R, Paninski L. Neuronal dynamics: from single neurons to networks and models of cognition. Cambridge, UK: Cambridge University Press; 2014.
34. Kiselev MV, Urusov AM, Ivanitsky AY. The adaptive Gaussian receptive fields for spiking encoding of numeric variables. Comput Res Model. 2025;17(3):389–400. doi:10.20537/2076-7633-2025-17-3-389-400.
35. Fang W, Yu Z, Chen Y, Huang T, Masquelier T, Tian Y. Deep residual learning in spiking neural networks. Adv Neural Inf Process Syst. 2021;34:21056–69.
36. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. Bull Math Biol. 1990;52(1–2):25–71. doi:10.1016/S0092-8240(05)80004-7.
37. Izhikevich EM. Simple model of spiking neurons. IEEE Trans Neural Netw. 2003;14(6):1569–72. doi:10.1109/TNN.2003.820440.
38. Abbott LF. Lapicque's introduction of the integrate-and-fire model neuron (1907). Brain Res Bull. 1999;50(5–6):303–4. doi:10.1016/s0361-9230(99)00161-6.
39. Wu Y, Deng L, Li G, Zhu J, Xie Y, Shi L. Direct training for spiking neural networks: faster, larger, better. Proc AAAI Conf Artif Intell. 2019;33(1):1311–8. doi:10.1609/aaai.v33i01.33011311.
40. Li B, Todo Y, Tang Z. Artificial visual system for orientation detection based on Hubel-Wiesel model. Brain Sci. 2022;12(4):470. doi:10.3390/brainsci12040470.
41. Cachi PG, Ventura S, Cios KJ. MT-SNN: spiking neural network that enables single-tasking of multiple tasks. arXiv:2208.01522. 2022.
42. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4510–20. doi:10.1109/CVPR.2018.00474.
43. Li K, Wan G, Cheng G, Meng L, Han J. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J Photogramm Remote Sens. 2020;159:296–307. doi:10.1016/j.isprsjprs.2019.11.023.
44. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, et al. MMDetection: open MMLab detection toolbox and benchmark. arXiv:1906.07155. 2019.
45. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, et al. SCRDet: towards more robust detection for small, cluttered and rotated objects. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 8231–40. doi:10.1109/ICCV.2019.00832.
46. Jiang Y, Zhu X, Wang X, Yang S, Li W, Wang H, et al. R2CNN: rotational region CNN for orientation robust scene text detection. arXiv:1706.09579. 2017.
47. Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, et al. Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans Multimedia. 2018;20(11):3111–22. doi:10.1109/tmm.2018.2818020.
48. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2999–3007. doi:10.1109/ICCV.2017.324.
49. Jiang Y, Tan Z, Wang J, Sun X, Lin M, Li H. GiraffeDet: a heavy-neck paradigm for object detection. arXiv:2202.04256. 2022.
50. Li Y, Wang H, Dang LM, Song HK, Moon H. ORCNN-X: attention-driven multiscale network for detecting small objects in complex aerial scenes. Remote Sens. 2023;15(14):3497. doi:10.3390/rs15143497.

51. Chen H, Chen K, Ding G, Han J, Lin Z, Liu L, et al. YOLOv10: real-time end-to-end object detection. In: Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 107984–8011. doi:10.52202/079017-3429.
52. Chen L, Zhang F, Xie G, Gao Y, Qi X, Sun M. S3Det: a fast object detector for remote sensing images based on artificial to spiking neural network conversion. Front Inform Technol Electron Eng. 2025;26(5):713–27. doi:10.1631/fitee.2400594.
53. Merolla PA, Arthur JV, Alvarez-Icaza R, Cassidy AS, Sawada J, Akopyan F, et al. Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science. 2014;345(6197):668–73. doi:10.1126/science.1254642.
54. Horowitz M. 1.1 computing's energy problem (and what we can do about it). In: 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC); 2014 Feb 9–13; San Francisco, CA, USA. p. 10–4. doi:10.1109/ISSCC.2014.6757323.
55. Ming Q, Zhou Z, Miao L, Zhang H, Li L. Dynamic anchor learning for arbitrary-oriented object detection. Proc AAAI Conf Artif Intell. 2021;35(3):2355–63. doi:10.1609/aaai.v35i3.16336.
56. Azimi SM, Vig E, Bahmanyar R, Körner M, Reinartz P. Towards multi-class object detection in unconstrained remote sensing imagery. In: Computer vision—ACCV 2018. Cham, Switzerland: Springer International Publishing; 2019. p. 150–65. doi:10.1007/978-3-030-20893-6_10.
57. Zhang G, Lu S, Zhang W. CAD-net: a context-aware detection network for objects in remote sensing imagery. IEEE Trans Geosci Remote Sens. 2019;57(12):10015–24. doi:10.1109/TGRS.2019.2930982.
58. Wu J, Pan Z, Lei B, Hu Y. FSANet: feature-and-spatial-aligned network for tiny object detection in remote sensing images. IEEE Trans Geosci Remote Sens. 2022;60:5630717. doi:10.1109/TGRS.2022.3205052.
59. Yang X, Yan J, Liao W, Yang X, Tang J, He T. SCRDet++: detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. IEEE Trans Pattern Anal Mach Intell. 2023;45(2):2384–99. doi:10.1109/TPAMI.2022.3166956.
60. Dong Y, Yang H, Liu S, Gao G, Li C. Optical remote sensing object detection based on background separation and small object compensation strategy. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;19:3341–51. doi:10.1109/JSTARS.2024.3351140.
61. Xu T, Sun X, Diao W, Zhao L, Fu K, Wang H. ASSD: feature aligned single-shot detection for multiscale objects in aerial imagery. IEEE Trans Geosci Remote Sensing. 2022;60:1–17. doi:10.1109/tgrs.2021.3089170.
62. Li Z, Wang Y, Zhang Y, Gao Y, Zhao Z, Feng H, et al. Context feature integration and balanced sampling strategy for small weak object detection in remote sensing imagery. IEEE Geosci Remote Sensing Lett. 2024;21:1–5. doi:10.1109/lgrs.2024.3356507.
63. Yu D, Ji S. A new spatial-oriented object detection framework for remote sensing images. IEEE Trans Geosci Remote Sens. 2022;60:4407416. doi:10.1109/TGRS.2021.3127232.
64. Gao T, Li Z, Wen Y, Chen T, Niu Q, Liu Z. Attention-free global multiscale fusion network for remote sensing object detection. IEEE Trans Geosci Remote Sens. 2024;62:5603214. doi:10.1109/TGRS.2023.3346041.
65. Zhou Z, Zhu Y. KLDet: detecting tiny objects in remote sensing images via kullback-leibler divergence. IEEE Trans Geosci Remote Sens. 2024;62:4703316. doi:10.1109/TGRS.2024.3382099.
66. Zhang T, Zhuang Y, Wang G, Dong S, Chen H, Li L. Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery. IEEE Trans Geosci Remote Sens. 2021;60:5608720. doi:10.1109/TGRS.2021.3108476.