

Article

Spiking Neural Network Based on Hierarchical Residual Quantization and Temporal Error Compensation for Remote Sensing Object Detection

Yaming Cao ^{1,2,3} , Yukai Xing ^{1,2,3}, Liqun Kuang ^{1,2,3} and Shichao Jiao ^{1,2,3,*} 

¹ The School of Computer Science and Technology, North China University, Taiyuan 030051, China; 20220087@nuc.edu.cn (Y.C.); x393039749@gmail.com (Y.X.); kuang@nuc.edu.cn (L.K.)

² Shanxi Key Laboratory of Machine Vision and Virtual Reality, Taiyuan 030051, China

³ Shanxi Vision Information Processing and Intelligent Robot Engineering Research Center, Taiyuan 030051, China

* Correspondence: 20230006@nuc.edu.cn

Abstract

Compared with traditional artificial neural networks (ANNs), spiking neural networks (SNNs) have lower computational complexity, lower energy consumption, and faster inference speed, making them more promising for practical deployment on edge devices. However, in SNNs, since the output spikes of neurons are discrete, the model may face information loss, especially when the membrane potential is quantized into binary spikes, where quantization errors can lead to model precision loss and information loss. To address these challenges, this study proposes a spiking neural network based on hierarchical residual quantization and temporal error compensation (HRQ-TEC-SNN) and used for remote sensing object detection tasks. Through hierarchical residual quantization and temporal error compensation design, higher resolution quantization of membrane potential can be performed and the quantization threshold of membrane potential can be dynamically adjusted to compensate for errors introduced during the quantization process, allowing for fine-tuning at each time step and reducing information loss caused by coarse quantization. In terms of network structure, by introducing depthwise separable convolution modules, channel attention and spatial attention mechanisms, and integrating fast spatial pyramid pooling based on pulse neural networks, the model's detection accuracy has been further improved while reducing the number of model parameters and computational costs. Experimental results show that the HRQ-TEC-SNN achieves significant advantages in both accuracy and energy consumption on the DOTAv1.0, DOTAv1.5 and DIOR datasets.

Keywords: spiking neural network; object detection; remote sensing image; hierarchical residual quantization; temporal error compensation; lightweight model



Academic Editor: Atsushi Mase

Received: 2 June 2026

Revised: 3 July 2026

Accepted: 6 July 2026

Published: 9 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Remote sensing object detection, as a core and critical task in the remote sensing data interpretation process, has irreplaceable application value in many national and livelihood fields such as land surveying, environmental monitoring, national defense security, agricultural yield estimation, urban planning, etc. [1]. Its detection accuracy, inference efficiency, and scene adaptation ability directly determine the conversion efficiency of remote sensing data from collection to application, and it is an important bridge connecting remote sensing technology with practical application needs [2]. With the rapid iteration and breakthrough

of deep learning technology in the field of computer vision, traditional remote sensing object detection methods based on manually designed features (such as HOG, SIFT, LBP, etc.) have gradually been replaced by detection methods based on deep neural networks due to their limited feature expression ability and insufficient generalization performance [3]. Among them, artificial neural networks (ANNs) and spiking neural networks (SNNs), as two representative deep learning models, have shown distinct technical advantages and broad development potential in the field of remote sensing object detection due to their unique structural advantages and technical characteristics [4].

In current academic research and engineering applications of remote sensing object detection, ANNs with their powerful hierarchical feature learning and adaptive feature extraction capabilities are an established a mature and complete technical system and application paradigm after years of technological iteration and scene adaptation, becoming the mainstream research method and practical deployment plan in this field [5–8]. From initially transferring mature pretrained models (such as AlexNet [9], VGG [10], ResNet [11], etc.) trained in the natural image field to remote sensing scenes, the significant superiority of deep features in characterizing remote sensing target details and distinguishing complex backgrounds compared to traditional manual features has been verified through fine-tuning and adaptation. Subsequently, specialized optimization and innovation have gradually been carried out to address the core problems of scale disparity, arbitrary direction, complex background, and dense targets unique to remote sensing targets, gradually forming a dedicated detection paradigm that is fully adapted to remote sensing scenes [12]. In the process of technological evolution, the optimization and innovation of detection frameworks have always been the core thread driving technological progress. From the gradual maturity and improvement of two-stage detection frameworks to the rapid rise and widespread application of one-stage detection models, from the limitations of traditional horizontal bounding box (HBB) detection that is difficult to adapt to targets in any orientation, to the breakthrough and popularization of oriented bounding box (OBB) technology, various improved models and key technologies (such as feature pyramid fusion, attention mechanism, anchor box optimization, etc.) continue to emerge, continuously improving the accuracy, efficiency, and robustness of remote sensing object detection [13,14]. At the same time, the continuous improvement and expansion of specialized datasets for remote sensing object detection (such as DOTA [15], HRSC2016 [16], etc.), as well as the innovative fusion of ANN and other deep learning technologies such as transformer [17], have further expanded the application boundaries of ANN in remote sensing object detection, enabling it to flexibly adapt to different types of remote sensing data such as synthetic aperture radar (SAR), hyperspectral, and multi-temporal data, as well as remote sensing detection needs in different scenarios such as ground, air, and satellite [18]. Although ANNs have shown many advantages in remote sensing object detection, they still have certain limitations in terms of high computational energy consumption, long inference time, and strong dependence on large-scale annotated data [19].

Unlike the mature development trend of ANNs, SNNs, as an emerging deep learning model that simulates the spike transmission mechanism of biological neurons, provide a new technological path for remote sensing target detection with its core characteristics of event-driven operation, low power consumption, and low latency [20,21]. In remote sensing detection scenarios, the widespread application of resource-limited platforms such as satellite and edge devices imposes strict requirements on the power consumption and latency of detection models. SNNs, through their unique method of encoding information with discrete spike sequences, reduce theoretical power consumption by at least one order of magnitude compared to traditional artificial neural networks [22], which perfectly

meets the application needs of such scenarios. Therefore, they have gradually become an emerging research hotspot in the field of remote sensing object detection.

However, compared to ANNs, the application of SNNs in complex visual tasks still faces many challenges. Particularly in remote sensing object detection tasks, the model not only needs to complete target category discrimination but also needs to accurately model continuous variables such as the spatial position, scale, and rotation angle of the target. The inherent discrete-time modeling and discrete pulse representation mechanism in SNNs makes the expression of continuous information highly dependent on statistics and aggregation in the time dimension, which can easily introduce quantization errors in the regression process and gradually accumulate in multi-time-step propagation, thereby adversely affecting detection accuracy [23]. In terms of SNN training, there is a lack of stable and efficient backpropagation (BP) training mechanisms like ANNs, which makes it difficult to train deep SNNs [24,25]. In addition, the existing SNN detection networks generally follow the design ideas of ANNs in structural design, and the problems of parameter redundancy and computational efficiency still exist, which to some extent weakens the energy efficiency advantage of SNNs.

In response to the above issues, this article focuses on the application of spiking neural networks (SNNs) in remote sensing object detection tasks and conducts research from two core levels: the neural mechanism and network structure and a spiking neural network based on hierarchical residual quantization and temporal error compensation (HRQ-TEC-SNN) is proposed. Firstly, the optimized design of hierarchical residual quantization neurons further refines the granularity of membrane potential quantification, improves quantification accuracy, and significantly reduces information loss caused by low-precision quantification. In the process of updating membrane potential, the temporal error compensation is explicitly introduced to compensate for the quantization result at the current time step, effectively suppressing the accumulation of quantization error in multi-time-step propagation and significantly improving the accuracy of membrane potential quantization and the modeling performance of SNN on continuous regression information. At the same time, in order to adapt to the characteristics of variable angles and disordered directions of remote sensing targets, HRQ-TEC-SNN integrates the Circular Smooth Label (CSL) [26] labeling strategy to achieve precise output of target rotation boxes. Secondly, based on the improvement of the neural mechanism, targeted optimization is carried out at the network structure level: lightweight convolution modules are introduced to streamline the model parameter scale, channel and spatial attention mechanisms are embedded to enhance effective feature screening, and a fast spatial pyramid pooling module based on SNNs is designed to improve feature extraction efficiency. While reducing model computational redundancy and improving energy efficiency, the spatial expression ability and feature selection ability of the model for remote sensing targets are also enhanced. Finally, a remote sensing rotating target detection SNN with synergistic optimization of surface energy efficiency and accuracy is constructed. The experimental results show that the proposed method further improves the accuracy of rotating object detection while significantly reducing model parameters and computational costs, effectively verifying the complementarity between temporal error compensation neuron modeling and network structure optimization strategies and providing a feasible solution for the efficient application of SNNs in remote sensing object detection.

The main contributions of this paper are threefold: (1) We propose a novel spiking neuron model that integrates hierarchical residual quantization (HRQ) and temporal error compensation (TEC), which effectively mitigates quantization errors and improves the representation of continuous regression information in SNNs. (2) We design a lightweight network architecture incorporating spiking depthwise separable convolution

(SNN_DSC), channel-spatial attention (CSAM), and an SNN-based fast spatial pyramid pooling (SNN_SPPF) to reduce computational cost while maintaining high detection accuracy. (3) We conduct extensive experiments on three public remote sensing datasets (DOTAv1.0 [15], DOTAv1.5 [27] and DIOR [28]) and demonstrate that our HRQ-TEC-SNN achieves superior accuracy–energy trade-offs compared to state-of-the-art ANN-based and SNN-based detectors.

The remainder of this paper is organized as follows. Section 2 reviews related work on remote sensing object detection, training methods for SNNs, and quantization errors. Section 3 presents the proposed HRQ-TEC-SNN in detail, covering hierarchical residual quantization, temporal error compensation, and the lightweight network modules. Section 4 describes the experimental setup, datasets, and evaluation metrics and reports the quantitative and ablation results. Section 5 discusses the synergistic effects of the modules. Finally, Section 6 concludes the paper and outlines future directions.

2. Related Work

This section reviews the most relevant prior research that forms the background of our work. We first discuss remote sensing object detection methods, with a focus on the differences between ANN-based and SNN-based approaches. Next, we survey existing training strategies for SNNs, including ANN-to-SNN conversion and direct training via surrogate gradients. Finally, we analyze the problem of quantization error in spiking neurons and summarize the limitations of current neuron models, which motivates our proposed design.

2.1. Remote Sensing Object Detection

With the rapid development of deep learning technology, remote sensing object detection methods have gradually formed a clear classification system, mainly divided into two categories: single-stage detectors and two-stage detectors. The two technical paths based on artificial neural networks (ANNs) and spiking neural networks (SNNs) have different application performances in remote sensing scenes due to their respective characteristics [29,30].

In the field of remote sensing object detection, breakthrough progress has been made in the technology route based on artificial neural networks (ANNs) [31]. In response to the problem of poor saliency of remote sensing target features, researchers use techniques such as prior knowledge of the target [32], adaptive adjustment of learning rate [33], and orthogonal strip convolution to make the model more stable in focusing on effective target information [34]. For the more common rotating targets in remote sensing scenes, such as ships, airplanes, and inclined bridges, the ANN method introduces targeted optimization design [35], oriented bounding box (OBB) or angle regression parameters [36], adaptive point learning methods [37], and rotation-equivariant features [38] to accurately describe the center coordinates, dimensions, and rotation angles of the rotating targets [39]. This effectively solves the problem of traditional axis-aligned bounding boxes being unable to adapt to the target's rotation posture and achieves accurate positioning and classification of rotating targets. However, such optimization designs often require additional angle regression branches and complex loss functions, resulting in a significant increase in the parameters and computational complexity of ANN models, greatly limiting the promotion and application of ANN in energy-sensitive remote sensing detection scenarios such as satellite and airborne applications [40].

Compared with an ANN, an SNN has the advantages of low energy consumption and high real-time performance by simulating the peak emission mechanism of biological neurons and has significant potential in the field of remote sensing object detection [41].

The existing SNN-based methods mostly adopt a single-stage detector framework, which can efficiently locate and classify remote sensing targets, adapt to resource-limited scenarios such as satellite and airborne applications, and are particularly good at processing remote sensing time-series data [42]. Due to the aforementioned advantages, event-driven SNN-based object detection methods have experienced rapid development across multiple dimensions, including datasets, network architectures, learning rules, and hardware deployment [43]. However, their application in remote sensing object detection has mostly been limited to horizontal targets, lacking adaptability to rotating targets, and related research remains in its early stages. Given the complex spatial distribution and attitude variations in rotating targets in remote sensing imagery, traditional spike encoding methods are difficult to adapt to SNN neuron structures, resulting in insufficient detection accuracy and robustness. At present, there have been preliminary explorations to combine SNN with rotation-invariant methods such as data augmentation [44] and angle parameterization [45] to improve performance, but further research is still needed.

2.2. Training Method of SNN

Due to the discrete nature of pulse signals, training spiking neural networks (SNNs) is challenging. Existing SNN training methods can primarily be categorized into two major approaches: ANN-to-SNN (ANN2SNN) conversion [23] and direct training of SNNs [42,46].

The ANN2SNN approach involves replacing the activation functions of traditional artificial neural networks with spiking neurons to achieve SNN training. Its core idea is to pretrain an artificial neural network (ANN) first, then migrate its parameters to an equivalent SNN model and perform inference through spike emission rules. This method leverages mature ANN training techniques and a wide range of optimized algorithms, avoiding the discrete signal optimization challenges encountered in direct SNN training. It also converges quickly, reducing training time. In order to further promote the engineering application of the ANN2SNN method in remote sensing scenes, Chen et al. proposed the S3Det model [47]. By introducing the Spiking Sequence Rapid Sensing (SSRS) module and Channel Self-decaying Weighted Normalization (CSWN) strategy, the S3Det model achieves efficient sampling and deep neuron activation control of remote sensing image pulse sequences. It also achieves detection performance similar to the original ANN model on datasets such as DOTA [15] and HRSC [48], with power consumption only 1/122 of the original ANN. However, ANN2SNN methods require long time steps to ensure precision, resulting in high inference latency, which does not align with the low-latency advantage of SNNs. Additionally, during the conversion process, the adjustment of membrane potential scales requires special quantization strategies; otherwise, it can lead to significant performance degradation. This can result in poor performance in complex tasks, particularly in dynamic scenarios or tasks like rotating object detection where requirements are challenging to meet [40].

Directly trained spiking neural networks (SNNs) address the issue of non-differentiable spike signals through the surrogate gradient [21] technique, enabling end-to-end training. Notable examples include the spatiotemporal backpropagation algorithm (STBP) [49] and the spike-based backpropagation algorithm [50]. The core idea involves simulating neuronal behavior during forward propagation using membrane potential accumulation and spiking emission rules and approximating the non-smooth aspects of spiking emission with surrogate gradient functions during backward propagation to facilitate gradient information transmission. The development of improved training techniques has promoted the exploration of object detection technologies based on spiking neural networks (SNNs). EMS-YOLO [46] adopts a fully spiking ResNet [11], which addresses the issues related to

non-spike-based computations found in previous models such as SEW-ResNet [51] and demonstrates robust performance in general object detection. Similarly, SpikeYOLO [42] introduces the I-LIF spiking neuron to convert non-spike-based computations into spike-based ones, thereby improving detection accuracy. However, these approaches are primarily designed for general object detection and perform suboptimally in remote sensing scenarios due to challenges such as diverse object angles and poor feature saliency.

2.3. Quantization Error

Quantization error represents a significant challenge for spiking neural networks (SNNs) in practical applications, primarily arising from the conversion of membrane potential signals into binary spikes. Traditional quantization methods often employ a coarse-grained strategy for membrane potentials, which simplifies the discretization process but leads to information loss, especially under low-precision quantization. This results in a diminished ability to accurately represent dynamic input signals. In the context of ANN-to-SNN conversion, information loss primarily stems from two sources: (1) the discrepancy between continuous ANN activations and discrete SNN spike rates and (2) the accumulation of quantization errors across time steps. Traditional conversion methods attempt to mitigate this by calibrating firing thresholds or employing rate-based approximations, but these approaches often require hundreds of time steps to achieve acceptable accuracy [52,53]. Additionally, the discontinuity inherent in fixed-threshold mechanisms, a crucial component of neuronal design, prevents precise alignment between membrane potential accumulation and spike generation with the dynamic characteristics of input signals, further contributing to quantization errors [54].

Current neuron designs still exhibit significant shortcomings in handling quantization errors. Traditional models, exemplified by the Leaky Integrate-and-Fire (LIF) [55] neuron, typically employ fixed thresholds and low-precision discretization schemes, resulting in poor adaptability when dealing with diverse inputs. On one hand, the fixed binarization mechanism struggles to capture subtle differences in input signals, particularly when input amplitudes are small, potentially failing to trigger spike generation and leading to information loss. On the other hand, the absence of an effective error feedback mechanism can cause quantization errors to accumulate progressively over time, thereby degrading overall performance. Additionally, the discrete-time decay process of membrane potential limits its responsiveness in scenarios involving rapid changes. Conventional quantization methods generally discretize numerical values with large step sizes, offering sufficient precision for most tasks while significantly accelerating model inference and reducing model size. However, when faced with complex tasks such as image recognition and speech recognition which demand high precision, coarse-grained quantization may fail to effectively capture minute differences in data, resulting in information loss and a decline in accuracy. This is particularly evident in the handling of fine-grained features or ambiguous edge regions, where coarse-grained quantization can lead to the loss of image details, consequently impacting recognition performance.

3. Methodology

In response to the issue of target information loss in remote sensing object detection tasks caused by membrane potential quantization errors in spiking neural networks, this study proposes HRQ-TEC-SNN for remote sensing object detection, whose core is integrated hierarchical residual quantization and temporal error compensation to mitigate quantization-induced errors and information loss, as shown in Figure 1. The upper left corner of Figure 1 shows an example of input–output, and below it is the structure of the HRQ-TEC-SNN model, consisting of a backbone and a detection head. The upper right

corner of Figure 1 shows how the various modules of the entire model are constructed. Below Figure 1 is the legend of the core module proposed in this study, which includes the full names of the corresponding modules. The HRC-TEQ module completes the optimization of neurons, while the SNN_DSC, SNN_SPPF, and CSAM modules jointly achieve a lightweight model structure. The image is repeatedly input at each time step T , and the first layer of spiking neurons is used to encode static images, serving as the input to the spiking neural network [51,56]. As the static image is encoded, the continuous pixel values in the image are gradually converted into spike signals, forming a series of time-varying spike events. This processing method not only enables the spiking neural network to capture image features through sequential spikes, effectively utilizing the spatial-temporal characteristics of SNNs, but also ensures that the input image information is fully propagated and processed in the temporal dimension, enhancing the model's sensitivity to spatial-temporal information.

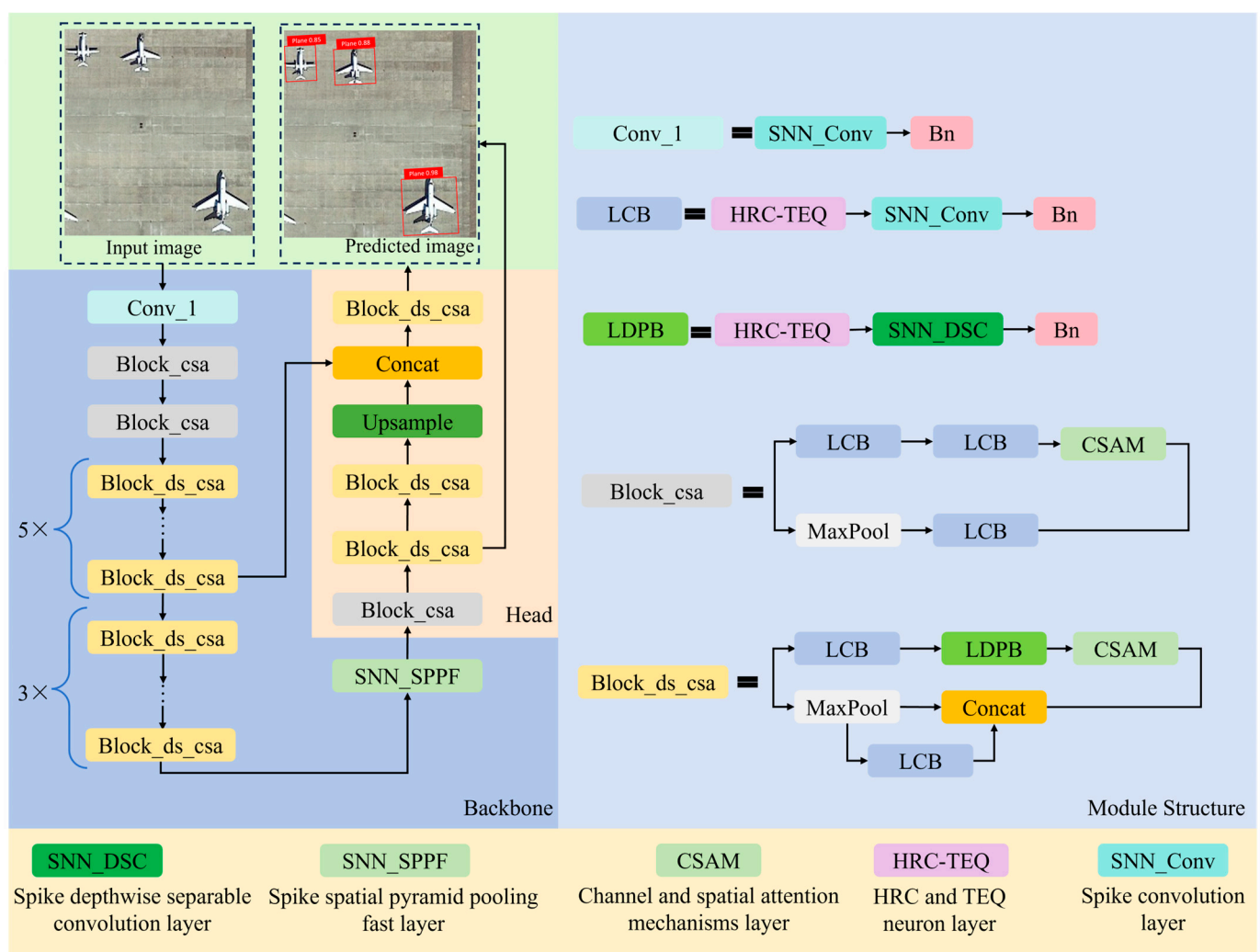


Figure 1. HRQ-TEC-SNN network architecture diagram, where “Bn” represents the batch normalization layer, “MaxPool” represents the max pooling layer, and “Upsample” represents the upsampling layer.

3.1. Hierarchical Residual Quantization

The hierarchical residual quantization (HRQ) module supports high-resolution processing of membrane potentials and mitigates information loss through fine-grained time-step optimization. The numerical quantification accuracy of membrane potentials at dis-

crete time steps is essential for maintaining the long-term stability and temporal continuity of error feedback mechanisms. Excessively large quantization errors will weaken or even eliminate residual information in the initial encoding phase, further leading to the failure of subsequent error feedback compensation. To address this issue, this work proposes a hierarchical residual quantization strategy. By narrowing the quantization step size, the proposed strategy enables precise characterization of the discretization of continuous data and effectively suppresses quantization errors throughout the encoding process.

Different from conventional coarse-grained quantization schemes, hierarchical residual quantization retains fine-grained data structural features more effectively. It is particularly applicable to high-precision downstream tasks and avoids performance degradation caused by excessive quantization deviation. For feature learning scenarios requiring subtle feature distinction, this quantization method substantially improves the representation precision of membrane potential dynamics and enhances the retention and transmission capability of error signals, thereby achieving prominent performance improvements in neural network models.

Specific quantitative formulas are as follows:

$$S(t) = Q(V(t)) = \text{round}\left(\frac{\text{clip}(V(t)), 0, V_{\max}}{\Delta v}\right) \cdot \Delta v \quad (1)$$

where $S(t)$ represents the spike output at time t ; $V(t)$ represents the membrane potential at time t ; $Q(\cdot)$ represents the multi-valued quantization function; $\text{clip}(\cdot)$ represents the truncation operation within the interval; $\text{round}(\cdot)$ represents rounding to the nearest discrete potential level; Δv represents the quantization granularity; and $V_{\max}$ represents the maximum quantizable membrane potential value. The hierarchical residual quantization strategy operates through a multi-level refinement process. At each quantization level l , the residual membrane potential $R_l(t)$ is computed as

$$R_l(t) = V(t) - \sum_{i=1}^l Q_i(V(t)) \quad (2)$$

where $Q_i(V(t)) = \text{round}(\text{clip}(V(t), 0, V_{\max}) / \Delta v_i) \cdot \Delta v_i$, and Δv_i denotes the quantization granularity at level i . The quantization granularity follows a hierarchical schedule $\Delta v_1 > \Delta v_2 > \dots > \Delta v_L$, where coarser levels capture the overall magnitude while finer levels encode residual details. The final quantized membrane potential is obtained as $S(t) = \sum_{l=1}^L Q_l(R_{l-1}(t))$, with $R_0(t) = V(t)$. This hierarchical design ensures that information not captured at coarser levels is preserved and propagated to finer levels, effectively mitigating the information loss typically associated with single-level coarse quantization.

The core superiority of hierarchical residual quantization stems from its refined quantization intervals, which efficiently alleviate information loss and retain the fine-grained details of original feature information. In contrast to conventional coarse-grained quantization schemes, this hierarchical residual paradigm enables more accurate characterization of network parameters and activation values, which substantially suppresses quantization errors and further strengthens the representational capability and inference precision of the model. As illustrated in Figure 2, distinct from the classic LIF neuron model, the proposed HRQ implements a granularity-g-dominated pulse-firing mechanism. Relying on the HRQ design, the dynamic variation range of membrane potential is constrained to a preset interval, and multi-level quantitative division is realized via discrete voltage levels $(0, g, 2g \dots)$. Once the membrane potential exceeds the corresponding threshold, neurons will fire spikes synchronously. This mechanism realizes direct quantitative mapping and effective representation of pulse outputs.

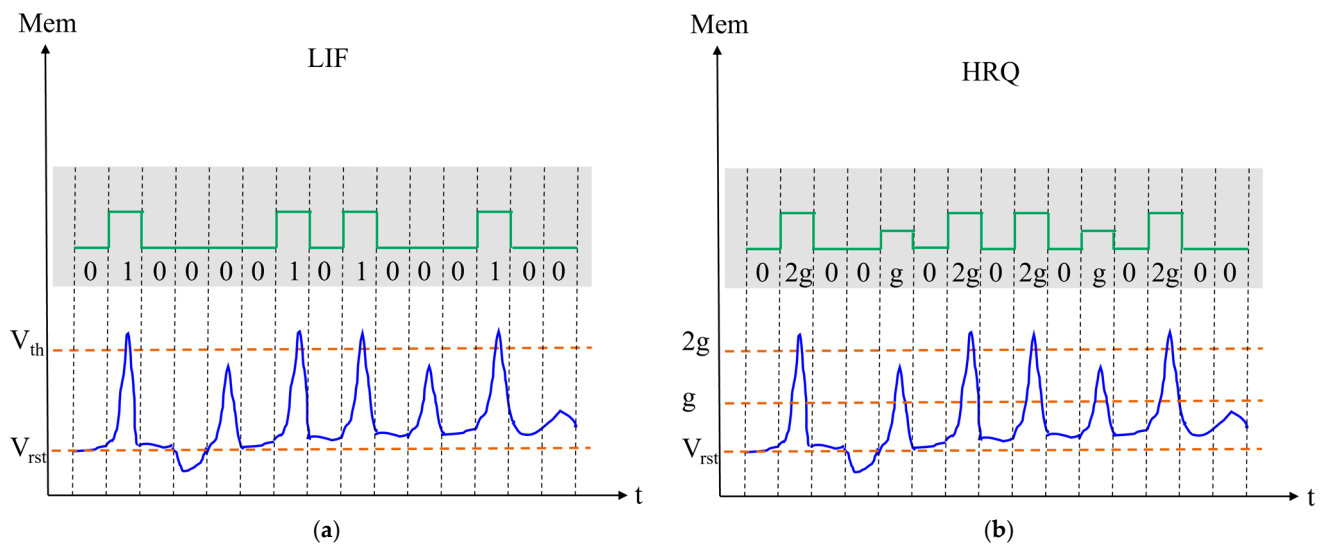


Figure 2. Comparison of activation potential, among them, the blue line represents the input analog voltage, the green line represents the pulse signal quantized by the pulse neuron of the analog voltage, the dashed line is an auxiliary illustration of discretizing the analog voltage with different thresholds, and the gray area is used to distinguish between the pulse signal and the analog voltage: (a) LIF neurons; (b) hierarchical residual quantization (HRQ).

It is important to clarify that the quantization thresholds are fixed during inference; they are determined by the granularity parameter g , which is optimized on the validation set. The proposed HRQ does not perform online adaptive adjustment of thresholds based on individual samples. Instead, it increases the number of quantization levels (from binary to multi-level) to reduce information loss, and the optimal g is selected based on the best trade-off between accuracy and inference speed (see Table 1). The proposed HRQ strategy addresses information loss from a fundamentally different perspective—rather than approximating ANN activations through spike rates, HRQ directly refines the quantization process of membrane potentials within the SNN itself. By employing multi-level quantization with progressively finer granularities, HRQ preserves fine-grained feature details that would otherwise be lost in conventional binary quantization. This is particularly beneficial for remote sensing object detection, where targets exhibit significant variations in scale, orientation, and appearance.

Table 1. Experimental results of different selection of threshold g on the DOTAv1.5 dataset.

g	mAP@0.5	mAP@0.5:0.95	FPS
0.3	66.7	47.1	37.88
0.5	67.2	47.3	42.55
0.7	64.8	46.7	37.45
0.9	63.1	46.1	40.65

In the hyperparameter selection phase, we conducted a fine-grained analysis of voltage threshold values of HRQ, with the experimental results presented in Table 1. These results indicate that the parameter g has a significant impact on the model's accuracy–speed trade-off. Specifically, as the threshold g increased from 0.3 to 0.9, the mAP@0.5 metric gradually declined from 66.7 to 63.1, while the mAP@0.5:0.95 metric decreased from 47.1 to 46.1. Notably, the model achieved its optimal accuracy metrics (both mAP@0.5 and mAP@0.5:0.95) when g was set to 0.5. In terms of inference efficiency, performance varied significantly across configurations. The model with g set to 0.5 demonstrated exceptional speed, reaching 42.55 FPS, which represents an improvement of 12.3%, 13.6%, and 4.7% over

the models with g set to 0.3, 0.7, and 0.9, respectively. A comprehensive analysis reveals that the configuration with g set to 0.5 yields the best overall performance, simultaneously enhancing accuracy and accelerating detection speed. Consequently, this study adopts a model with g set to 0.5 as the deployment solution to achieve both high precision and real-time performance in practical environments.

To investigate the impact of the number of quantization levels on model performance, we conducted experiments with varying numbers of HRQ levels ($L = 2, 4, 6, 8$) on the DOTA1.5 dataset. The results are summarized in Table 2. The results indicate that increasing the number of quantization levels from 2 to 4 yields substantial accuracy improvements (+1.1% mAP@0.5) with modest increases in energy consumption (+0.6 mJ) and inference time (+1.8 ms). However, further increasing to 6 or 8 levels provides diminishing returns in accuracy while incurring progressively larger computational overheads. Based on this analysis, we select $L = 4$ as the optimal configuration, balancing detection accuracy with computational efficiency.

Table 2. Experimental results of impact of quantization levels on the DOTA1.5 dataset.

Levels (L)	mAP@0.5	mAP@0.5:0.95	Energy (mJ)	Inference Time (ms)
2	65.8	36.9	18.7	22.3
4	66.9	46.5	19.3	24.1
6	66.7	46.1	22.4	26.8
8	66.3	46.2	25.8	30.5

3.2. Temporal Error Compensation

To improve detection stability and precision, this paper proposes a temporal error compensation (TEC) spiking neuron module. During forward propagation, the module temporally quantifies the deviation between membrane potential and spike outputs, and further optimizes membrane potential updating to achieve more accurate neuronal spike firing. Built upon the conventional LIF neuron framework, the TEC module takes the discrepancy between instantaneous membrane potential and spike output as quantization error. Such error is subsequently fed back to the subsequent quantization iteration for progressive membrane potential calibration, which facilitates high-fidelity spike generation. The proposed module embeds a dynamic compensation mechanism into the spike generation procedure. Specifically, information loss induced by incomplete spike encoding is remedied through residual components embedded in membrane potential updates, which are delivered to subsequent time steps. This establishes a complete temporal residual compensation loop, enabling neurons to compensate for inadequate spike firing at the current moment in subsequent temporal iterations. Experimental results demonstrate that the TEC module can effectively mitigate performance degradation stemming from quantization distortion and information attenuation. Meanwhile, it greatly enhances the network's capability in capturing target boundaries, orientation angles and spatial positions, and ultimately improves the overall recognition accuracy. The temporal error at time step t is formally defined as the quantization residual:

$$E(t) = V(t) - S(t) \quad (3)$$

where $V(t)$ is the membrane potential and $S(t)$ is the quantized spike output. This error represents the information lost during the spike generation process—the portion of membrane potential that failed to be expressed as spikes at the current time step.

The TEC module incorporates this error into the subsequent time step's membrane potential update:

$$V(t+1) = V(t) + \frac{\Delta t}{\tau} [-V(t) + RI(t)] + E(t) \quad (4)$$

Let Δt be the discrete time step and τ be the membrane time constant; their combination $\frac{\Delta t}{\tau}$ forms the leakage term. Let $V(t)$ represent the membrane potential, $I(t)$ represent the input current, $E(t)$ represent the error feedback term, and $S(t)$ represent the spike output at time t . Whenever a neuron receives an input signal, the membrane potential updates based on the previous state, the current input signal, and the spike output from the previous step.

This formulation can be derived by considering the ideal membrane potential dynamics $\tilde{V}(t+1) = V(t) + \frac{\Delta t}{\tau} [-V(t) + RI(t)]$ and the quantization error introduced at time t . By adding $E(t)$ to the update, we effectively compensate for the information that was missed in the previous spike generation.

Importantly, the error feedback establishes a forward compensation loop: $E(t)$ influences future membrane potentials but does not receive gradients during backpropagation. This is achieved by treating $E(t)$ as a constant in the computational graph, ensuring that gradient propagation follows the standard LIF path through the surrogate gradient function. The stability of this compensation is inherently guaranteed as the quantization function bounds $E(t)$ within $[0, \Delta v]$, preventing uncontrolled error accumulation.

Critically, the gradients are not backpropagated through $E(t)$. The direct current estimator performs backpropagation on $S(t)$ after Equation (2), while $E(t)$ serves as a constant that only participates in forward computation for subsequent time steps. The gradient propagates along the path from $V(t)$ to $S(t)$ through the differentiable agent function, which is completely consistent with the traditional LIF neuron, ensuring the feasibility of gradient calculation and training stability. $E(t)$ establishes a pure forward compensation loop that remains unaffected by gradient computation. The stability of the forward compensation is inherently guaranteed as the quantization function $Q(\cdot)$ bounds the values of $E(t)$, preventing uncontrolled growth.

As shown in Figure 3, two consecutive time steps are used to illustrate the update mechanism of TEC. The TEC mechanism is achieved by integrating the quantization error $E(t-1)$ at time $t-1$ into the membrane potential $Spike(t)$ process at time t . In contrast to traditional LIF neurons, the TEC module exhibits stronger self-regulation capabilities. Traditional LIF neurons are static and memoryless; once an input fails to trigger a spike, that information is permanently lost. However, TEC neurons are dynamic and adaptive, quantifying the unexpressed portion as an error and incorporating it into subsequent membrane potential evolution. Information not expressed during forward propagation does not vanish but is preserved over time and reintegrated at subsequent time steps, thereby enhancing the network's expressive power and robustness. The TEC mechanism fundamentally alters spike generation patterns by introducing a temporal compensation loop that addresses quantization errors across multiple dimensions. By compensating for quantization errors from previous time steps, TEC enables neurons to generate spikes that more accurately reflect the temporal evolution of input signals, with error compensation smoothing out the discontinuities introduced by binary quantization and resulting in more consistent firing rates across time steps compared to conventional LIF neurons. Furthermore, TEC enhances the temporal fidelity of spike patterns by preserving information that would otherwise be permanently lost in conventional LIF neurons when it fails to trigger a spike at time t ; with TEC, this information is retained as $E(t)$ and carried forward, allowing the neuron to “remember” and compensate for missed activations, which is particularly beneficial for detecting small or partially occluded targets in remote

sensing imagery where weak signals may not exceed the firing threshold in a single time step but become detectable when accumulated over multiple steps. As a result, TEC improves feature representation quality by reducing the distortion introduced by quantization, thereby enabling more accurate and robust feature learning for complex detection tasks.

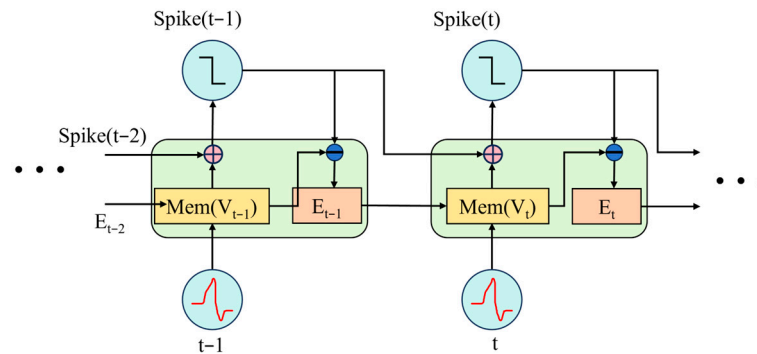


Figure 3. Schematic diagram of temporal error compensation (TEC) mechanism.

Our analysis reveals that the TEC module introduces minimal computational cost. Specifically, the error compensation operation requires only one additional addition per neuron per time step to store and incorporate the error term $E(t)$ during membrane potential updates, which translates to approximately 3.8% additional FLOPs compared to the baseline LIF-based model. Importantly, the TEC mechanism introduces no additional learnable parameters, as it solely modifies the neuron dynamics without adding extra weight layers or convolution operations. In terms of inference speed, the TEC-equipped model increases inference time by approximately 3.9% (from 23.2 ms to 24.1 ms per image), while energy consumption rises marginally from 19.4 mJ to 20.1 mJ (+3.6%).

Balanced against these modest overheads, the TEC mechanism yields substantial accuracy improvements. When combined with HRQ, TEC improves mAP@0.5 from 64.7% to 67.2% on the DOTA v1.5 dataset as Section 4 shows, representing a gain of 2.5 percentage points. This translates to an efficiency ratio of approximately 0.65 percentage points of mAP improvement per 1% of additional computational cost—a highly favorable trade-off compared to typical accuracy–efficiency trade-offs in deep learning models. The overhead-to-gain ratio is particularly compelling when considering that the additional computation is purely additive (accumulation operations at 0.9 pJ) rather than multiplicative (multiply–accumulate at 4.6 pJ), aligning well with the energy-efficient nature of SNN computation. In summary, while the TEC mechanism introduces a modest computational overhead, the resulting accuracy gains substantially outweigh this cost. For latency-critical deployment scenarios, users may optionally reduce the number of time steps or quantization levels to mitigate this overhead while still benefiting from the temporal error compensation effect.

During hierarchical residual quantization with temporal error compensation, the proxy gradient method is employed to address the non-differentiability of gradients caused by quantization errors, enabling effective training of spiking neural networks. By using a rectangular window as the proxy gradient during model training, approximate gradient information can be provided for the quantization operations, which facilitates effective backpropagation and parameter updates, thereby completing the training process. The rectangular window is defined in Equation (4):

$$\frac{\partial S(t)}{\partial V(t)} = \begin{cases} 1, & 0 \leq V(t) \leq V_{max} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The rectangular window method serves as a straightforward yet robust technique for gradient approximation. This approach addresses the non-differentiable nature of quantization by assigning a constant gradient within each quantization interval, effectively modeling the gradient surrogate as a rectangular shape. Such a formulation allows the impact of quantization on gradients to be linearly approximated during backpropagation, thereby safeguarding the stability and convergence of the model throughout training. By utilizing this rectangular window as a gradient surrogate, the method facilitates effective gradient propagation while maintaining fine-grained quantization precision. Consequently, it ensures that the model can be successfully optimized under quantization constraints, yielding superior performance in complex tasks.

We further investigated the impact of simulation time steps T on detection performance. Experiments were conducted with $T = 2, 4, 6, 8, 10$ on the DOTAv1.5 dataset, with results shown in Table 3. The results demonstrate that detection accuracy improves rapidly as T increases from 2 to 4, after which the gains become marginal. This suggests that the HRQ-TEC mechanism effectively compensates for temporal information loss even with relatively few time steps. This efficiency advantage is attributed to TEC's ability to preserve and propagate information across time steps, reducing the need for extensive temporal accumulation. For practical deployment, we recommend $T = 4$ as the optimal choice, offering the best trade-off between accuracy and computational efficiency.

Table 3. Experimental results of impact of simulation time steps on the DOTAv1.5 dataset.

Time Steps (T)	mAP@0.5	mAP@0.5:0.95	Energy (mJ)	FPS
2	62.3	35.1	10.5	68.2
4	68.4	47.8	19.9	42.6
6	68.5	44.2	32.6	30.1
8	68.3	46.2	45.3	23.5
10	68.1	39.3	50.5	19.2

3.3. Lightweight Network Design for Energy Efficiency Precision Collaboration

The HRQ and TEC modules proposed in this article effectively optimize the neurons in the SNN by refining the membrane potential quantization threshold and introducing quantization error compensation, effectively alleviating the error accumulation problem generated during the discretization process of SNNs. However, when relying solely on optimization at the neuron level it is still difficult to fully unleash the potential of SNNs in terms of energy efficiency.

The existing network structure generally suffers from problems such as a large parameter count, redundant computation, and insufficient multi-scale feature expression ability. This not only leads to increased inference latency but also offsets some energy consumption advantages in actual deployment. Therefore, in order to further improve the deployment efficiency of the model at the edge, this paper proposes a lightweight network design for energy efficiency and accuracy collaborative optimization based on the improvement of the neuron mechanism. Specifically, this article introduces three complementary optimization strategies at the network structure level. Firstly, depthwise separable convolution (DS) is adopted to replace some traditional standard convolutions, in order to reduce the parameter size and computational redundancy of the model, thereby improving the inference efficiency of the pulse network on edge devices. Secondly, channel attention (CA) and spatial attention (SA) modules are introduced to enhance the network's spatial description and feature selection ability for key regions of rotating targets, improving target discriminability in complex scenes. Finally, by integrating the SNN-Spatial Pyramid Pooling Fast (SNN_SPPF) module based on spiking neural networks, the multi-scale feature fusion

capability is enhanced, enabling the network to fully capture the structural information of targets at different scales.

3.3.1. Spike Depthwise Separable Convolution Layer

Standard convolution has strong expressive power, being especially suitable for the front-end and prediction stages of networks that require fine-grained feature modeling and high-resolution texture preservation. However, its computational cost is high, and it requires strict storage and computing resources. In the deep feature abstraction stage, it is prone to generate a large number of redundant operations, which is not conducive to edge-device deployment and real-time inference scenarios. Therefore, this study introduces depthwise separable convolution (DS) as a replacement for some standard convolution structures to improve the overall computational efficiency of the network [57]. If the input feature size is $H \times W$, the input channel number is C_{in} , the output channel number is C_{out} , and the convolution kernel size is $K \times K$, then the parameter and computational complexity of the standard convolution are:

$$Params_{std} = K^2 \cdot C_{in} \cdot C_{out} \quad (6)$$

$$FLOPs_{std} = K^2 \cdot C_{in} \cdot C_{out} \cdot H \cdot W \quad (7)$$

Unlike standard convolution, DS divides convolution operations into two stages: Depthwise Convolution and Pointwise Convolution. The former independently applies $K \times K$ convolution to each input channel to extract spatial structure information, while the latter achieves cross-channel information fusion through 1×1 convolution to achieve feature compression and structural sparsity. The number of parameters and computational complexity are as follows:

$$Params_{DS} = K^2 \cdot C_{in} + C_{in} \cdot C_{out} \quad (8)$$

$$FLOPs_{DS} = K^2 \cdot C_{in} \cdot H \cdot W + C_{in} \cdot C_{out} \cdot H \cdot W \quad (9)$$

The ratio of the two is

$$\frac{Params_{DS}}{Params_{std}} = \frac{1}{C_{out}} + \frac{1}{K^2} \quad (10)$$

When $K = 3$ and $C_{out} = 64$, it can be reduced to about 1/8 of the standard convolution, and the number of parameters is significantly reduced. Due to the consistency of the input and output feature dimensions, DS only changes the feature transformation method without compressing the expression space. Therefore, while significantly reducing redundant calculations and parameter size, it can still maintain the modeling ability of key features. Therefore, depthwise separable convolution has natural advantages in lightweight networks and can improve inference speed, reduce storage overhead, and enhance deployment friendliness, making it particularly suitable for mobile and edge-device scenarios.

In this study, the DS global replacement network was not fully convolved, but was selectively deployed only in the deep regions of the backbone and the head downsampling stage, while retaining the standard convolution structure in the front section of the backbone and the rest of the detection head. Its design is based on the differences in network hierarchy functions: the backbone is responsible for advanced semantic abstraction in the deep layer, and the feature map space size is already small, while the channel dimension is large. Standard convolution has the highest cost and most redundant features in this area, so using DS can significantly reduce computational overhead while ensuring semantic expression ability; in the shallow and head stages of backbone, the network relies more on high-resolution textures and spatial details to support final localization and classification,

so retaining standard convolutions helps avoid information loss. This selective replacement hybrid design effectively balances structural lightweighting and accuracy preservation, allowing the model to maintain feature representation while reducing parameter count and inference latency.

3.3.2. Channel and Spatial Attention Mechanism Layer

To further enhance the spatial localization ability and structural detail expression ability of the model for rotating targets, this paper introduces channel attention (CA) and spatial attention (SA) mechanisms [58] to HRQ-TEC-SNN to achieve dual feature enhancement in both semantic and spatial dimensions. Channel attention (CA) models the global semantic correlation between different channels, adaptively adjusts channel weights, and focuses on key semantic channels related to the target while suppressing redundant background information. Spatial attention (SA) generates spatial saliency weights by aggregating the feature maps of channels, enabling the model to highlight the structure of the target region and form stronger discriminative ability in terms of rotation angle, edge contour, and local salient region expression. The attention combination module is inserted into the back half of the model and the key nodes in front of the head to enhance the discriminative ability of highly semantic features, especially in scenes with significant differences in rotating target scales and complex backgrounds, which can effectively improve the problem of feature confusion.

3.3.3. Spike Spatial Pyramid Pooling Fast Layer

In object detection tasks, the shape, structural texture, and spatial response range of targets at different scales vary significantly, and multi-scale feature expression has a crucial impact on detection performance. Therefore, this article further designs a multi-scale fusion module SNN_SPPF suitable for pulse neural networks based on SPPF (Spatial Pyramid Pooling Fast). Unlike the continuous feature flow in a traditional ANN, the SNN adopts a membrane potential accumulation and pulse issuance mechanism, so it is necessary to consider the computational consistency and pulse sparsity properties in the time dimension. SNN_SPPF introduces pulse convolution and membrane potential update functions into the SPPF structural framework, ensuring that the feature construction process before and after pooling conforms to the pulse-based dynamic accumulation law, and compresses and fuses the channels through 1×1 convolution to reduce the computational overhead and storage pressure during the pooling process, while avoiding overly dense pulse distribution. Subsequently, the features are subjected to three maximum pooling operations in sequence, and the pooling results of different scales are concatenated into channels. Finally, the pulse convolution fusion is used to form a unified pulse output, achieving spatiotemporal joint feature enhancement. The formula for SNN_SPPF is as follows:

$$X'_t = \text{Conv}_{snn}(\text{MU}(X_t)) \quad (11)$$

$$Y_t = \text{Conv}_{snn}(X'_t, \text{MP}_k(X'_t), \text{MP}_k(\text{MP}_k(X'_t)), \text{MP}_k(\text{MP}_k(\text{MP}_k(X'_t)))) \quad (12)$$

Among them, X_t represents the input pulse characteristics at time step t ; $\text{MU}(\cdot)$ is the membrane potential update function used to achieve pulse accumulation and discharge; $\text{Conv}_{snn}(\cdot)$ represents a convolution operator based on pulse neurons; X'_t is the feature after pulse preprocessing; $\text{MP}_k(\cdot)$ represents the maximum pooling operation with a kernel size of k and a stride of 1; $[\cdot]$ represents the stitching operation along the channel dimension; and Y_t is the multi-scale fusion feature output by the SNN_SPPF module. By introducing SNN_SPPF, this network can adapt to the temporal computing mechanism of SNNs while maintaining the fast multi-scale aggregation advantage of SPPF, effectively enhancing

the structural expression ability of targets of different scales and rotation angles, and supporting more efficient edge deployment on the basis of lightweight optimization.

4. Experiments

4.1. Dataset and Evaluation

This paper conducts experiments using the DOTA v1.0 [15], DOTA v1.5 [27] and DIOR datasets to validate the effectiveness of the HRQ-TEC-SNN model in remote sensing object detection tasks.

DIOR Dataset: The DIOR dataset comprises 23,463 optical remote sensing images with a resolution of 800×800 pixels. It contains 192,472 annotated instances across 20 object categories: airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf course, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. The dataset is split into training (11,725 images), validation (5863 images), and testing (5875 images) sets following the official protocol. The images are collected from Google Earth and have a spatial resolution ranging from 0.5 to 2 m per pixel, covering diverse scenes including urban, rural, and coastal areas.

DOTA v1.0 Dataset: The DOTA v1.0 dataset contains 2806 high-resolution aerial images (ranging from 800×800 to 4000×4000 pixels) with 188,282 annotated instances across 15 categories: plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer field, and swimming pool. Following the standard protocol, the dataset is divided into training (1411 images), validation (458 images), and testing (937 images) sets. All images are cropped into 1024×1024 patches with an overlap of 200 pixels during training and testing.

DOTA v1.5 Dataset: DOTA v1.5 extends DOTA v1.0 by adding annotations for smaller objects (minimum size reduced from 10 to 5 pixels) and introducing a new category “container crane”. The total number of instances increases to approximately 400,000, making small-object detection significantly more challenging. The image resolution and split protocol remain consistent with DOTA v1.0.

All models were trained on an NVIDIA GeForce RTX 4090 Ti GPU with 24 GB memory (produced in Asustek, Taiwan, China). We employed the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.937 and weight decay of 0.0005. The initial learning rate was set to 0.01 and scheduled using a cosine annealing decay with warm-up. Specifically, we used a linear warm-up for the first 3 epochs, increasing the learning rate from 0.001 to 0.01, followed by cosine decay to 0.0001 over the remaining epochs. The models were trained for 300 epochs with a batch size of 16. Input images were resized to 640×640 pixels during training. Data augmentation included random horizontal flipping (probability 0.5), random rotation ($\pm 90^\circ$), random scaling (0.5 to 1.5), and mosaic augmentation (applied for the first 100 epochs). For the DOTA datasets, we employed the standard cropping strategy with a crop size of 1024×1024 and overlap of 200 pixels. For ANNs (E_{ANN}) and SNNs (E_{SNN}), the specific calculation methods are as Equations (12) and (13).

$$E_{ANN} = O^2 \times C_{in} \times C_{out} \times K^2 \times E_{MAC} \quad (13)$$

$$E_{SNN} = T \times (fr \times E_{AC} \times OP_{AC} + E_{MAC} \times OP_{MAC}) \quad (14)$$

where T represents the total time step, O represents the feature output size, and C_{in} and C_{out} represent the number of input and output channels, respectively. K is the kernel size, and fr represents the average spike firing rate. OP_{AC} and OP_{MAC} denote the number of AC (accumulate) and MAC (multiply-accumulate) operations, respectively. In energy

consumption calculations, the most used energy consumption evaluation methods for SNNs are followed [59]. With all operations implemented based on 32-bit floating-point arithmetic under a 45-nanometer process, where E_{MAC} is 4.6 pJ, and E_{AC} is 0.9 pJ [60].

To comprehensively evaluate the detection performance of the model on all object categories within the dataset, this paper employs mAP@0.5 and mAP@0.5:0.95 as evaluation metrics for the object detection task. MAP (mean Average Precision) is a widely used metric for evaluating the performance of object detection tasks, particularly in the field of computer vision, where it provides a comprehensive assessment of detection precision and recall for all object categories in an image. By comparing detection results with ground-truth annotations, mAP helps quantify the performance of the algorithm. The evaluation metrics for model energy consumption adopt the methods proposed on the DOTA v1.5 dataset. Furthermore, to holistically assess the model's computational efficiency and practicality for deployment on resource-constrained platforms, we additionally adopt FLOPs (Floating-Point Operations) as key metrics for evaluating algorithmic complexity. The time complexity of a traditional ANN is mainly based on the network topology structure, while the calculation of SNNs is unfolded in time and space. When directly using FLOPs to measure the time complexity of SNNs, it only reflects the worst-case scenario of SNN theoretical peak calculation requirements, and it is difficult to reflect the actual operating efficiency and potential of an SNN. Therefore, in response to the time complexity of SNNs, this article adopts SOPS (synaptic operations per second) to calculate the actual operational efficiency and potential of the SNN [61]. The SOPS calculation of the pulse neural network considers the convolution operation of each layer and the pulse emission of the previous layer, as well as the floating-point addition involved in neuron charging. By adding the pulse-firing rate of each layer of the network to the calculation process, the corresponding SOPS result of the pulse neural network can be accurately obtained. The specific calculation formula is as follows:

$$SOPS = \sum_{t=1}^T \sum_{i=1}^N ((f_{i-1,t} \times C_{in,i} \times k_{h,i} \times k_{w,i} + 1) \times C_{out,i} \times H_{out,i} \times W_{out,i}) \quad (15)$$

Among them, T represents the total time step, N represents the total number of layers in the pulse neural network, $f_{i-1,t}$ represents the pulse-firing rate of neurons in the $i - 1$ layer at time t , $C_{in,i}$ and $C_{out,i}$ represent the number of input and output channels of the i -th convolutional layer, $k_{h,i}$ and $k_{w,i}$ represent the size of the i -th convolutional kernel, and $H_{out,i}$ and $W_{out,i}$ represent the output dimension of the i -th feature map.

4.2. Quantitative Experimental Comparison

As shown in Table 4, comparative experiments on the DOTA v1.0 dataset demonstrate that our method achieved an mAP50 of 78.40 with a lightweight configuration of only 19.4 M parameters, demonstrating significant advantages in the trade-off between accuracy and parameters. Compared with classic rotation object detection methods such as ReDet (31.6 M, 76.25) and O-RepPoints (36.6 M, 75.97), our model achieves higher or equivalent detection accuracy while significantly reducing the number of parameters. Compared with the recent high-performance models PKINet (30.8 M, 78.39) and RTMDet (52.3 M, 78.85), our method achieved the same level of absolute accuracy while reducing the parameter size by about 37% and 63%, respectively, demonstrating clear advantages of lightweight operation and deployment friendliness. In addition, compared with traditional single-stage methods such as S2ANet, R3Det, etc., the overall performance of our method is in the leading tier, significantly improving the competitiveness of the single-stage rotation detection framework in terms of accuracy.

Table 4. Experimental results of different models on the DOTA v1.0 dataset.

	Models	Params (M)	mAP@0.5	FLOPs (G)/SOPS	Energy (mJ)
Two-Stage	CenterMap [62]	41.1	71.59	198	455.4
	SCRDet [63]	41.9	72.61	-	-
	RoI Trans [13]	55.1	74.05	200	460.0
	FR-Est [64]	-	74.20	-	-
	G.Vertex [65]	41.1	75.02	198	455.4
	ARC [66]	74.4	77.35	212	487.6
	LSKNet-S [32]	31.0	77.49	161	370.3
	PKINet [67]	30.8	78.39	-	-
One-Stage	R3Det [45]	41.9	69.70	336	772.8
	S2Anet [68]	38.5	74.12	198	455.4
	O-RepPoints [37]	36.6	75.97	-	-
	R3Det-GWD	41.9	76.34	336	772.8
	R3Det-KLD	41.9	77.36	336	772.8
	RTMDet [69]	52.3	78.85	205	471.6
	HRQ-TEC-SNN (ours)	19.4	78.40	104.2	20.1

The experimental results on the DOTA 1.5 dataset are shown in Table 5. As indicated by the data in Table 5, HRQ-TEC-SNN demonstrates a clear energy efficiency advantage while maintaining high detection accuracy (mAP@0.5 = 73.70%), with an energy consumption of only 20.1 mJ, which is 8% and 3% of ReDet and LSKNet, respectively. This highlights the efficiency and sparsity characteristics of SNNs under event-driven computation. By analyzing the results based on the types of pulse neural networks, it can be seen that our method has certain advantages in detection accuracy (mAP@0.5), energy consumption (Energy), and Params compared to spiking transformers (Meta-SpikeFormer) and hybrid neuromorphic models (EMS-Res-SNN).

Table 5. Experimental results of different models on the DOTA v1.5 dataset.

	Models	Params (M)	mAP@0.5	Flops (G)/SOPS	Energy (mJ)
	AO2-DETR13 [70]	39.8	66.26	148.0	680.8
	DCFL [71]	36.1	66.80	157.7	725.4
	RoI Trans [13]	55.1	65.50	200	460.0
	ReDet [38]	31.6	66.86	108.8	250.4
	LSKNet [32]	31.0	70.26	160.7	739.2
	Meta-SpikeFormer [18]	28.6	46.00	-	74.8
	EMS-Res-SNN [46]	34.3	56.00	127.4	72.3
	O-RepPoints [37]	36.6	66.90	-	-
	SpikeYOLO [42]	68.8	65.70	144.8	84.2
	RTMDet [69]	52.3	71.80	205	471.6
	HRQ-TEC-SNN (ours)	19.4	73.70	104.2	20.1

It also can be seen that the HRQ-TEC-SNN model proposed in this paper has lower computational complexity compared to other models. This gives the HRQ-TEC-SNN model certain advantages in practical deployment. All the model indicators in Table 5 fully validate the practical potential and deployment value of HRQ-TEC-SNN in high-efficiency remote sensing scenarios. A detailed comparison with state-of-the-art SNN-based detectors reveals that HRQ-TEC-SNN achieves superior accuracy–energy trade-offs. For instance, on DOTAv1.5, our method obtains 73.7% mAP@0.5 with only 20.1 mJ, whereas SpikeYOLO reaches 65.7% at 84.2 mJ, EMS-Res-SNN achieves 56.0% at 72.3 mJ, and Meta-SpikeFormer attains only 46.0% at 74.8 mJ. Against ANN-based oriented detectors, our model not only achieves comparable or higher accuracy (e.g., 74.7% on DIOR vs. 67.03% for PKINet and 68.19% for MessDet) but also uses far fewer parameters (19.4 M vs. 30.8 M for PKINet and 52.3 M for RTMDet). In terms of energy, our consumption is only 1/23 of ReDet and 1/37 of LSKNet on DOTAv1.5, clearly demonstrating the advantage of event-driven computation.

As shown in Table 6, the comparative experiments on the DIOR dataset further validated the superiority and generalization ability of HRQ-TEC-SNN in rotating object detection tasks. Compared with the classic single-stage and two-stage rotation detection baselines, the HRQ-TEC-SNN proposed in this paper achieves optimal detection accuracy while maintaining an extremely low number of parameters. Specifically, traditional methods such as RetinaNet OBB (36.4 M, 57.55), Faster R-CNN OBB (41.1 M, 59.54), and RoI Transformer (55.1 M, 63.87) have significantly lower accuracy than our proposed method, indicating that traditional rotation detection architectures lack sufficient expressive power when facing more complex background and target deformations in DIOR. The experimental results across the three datasets also reveal the strengths and weaknesses of our approach. On DOTAv1.0 and DIOR, which contain relatively larger objects and diverse categories, our model achieves state-of-the-art or competitive accuracy, demonstrating strong generalization. In contrast, on DOTAv1.5—which includes many small instances and dense scenes—the performance drops slightly (73.7% vs. 78.4% on DOTAv1.0). This indicates that the sparse spike representation in SNNs may not yet be optimal for extremely small targets, and improving small-object detection remains an important direction for future work.

Table 6. Experimental results of different models on the DIOR dataset.

Models	Params (M)	mAP@0.5
RetinaNet OBB [72]	36.4	57.55
Faster R-CNN OBB [73]	41.1	59.54
RoI Transformer [13]	55.1	63.87
LSKNet [32]	31.0	65.90
O-RepPoints [37]	36.6	66.71
DCFL [71]	-	66.80
RTMDet [69]	52.3	66.86
PKINet [67]	30.8	67.03
MessDet [74]	18.1	68.19
HRQ-TEC-SNN (ours)	19.4	74.7

Compared with more advanced lightweight or rotation-task-optimized models in recent times, HRQ-TEC-SNN also demonstrates significant advantages. For example, in comparison with LSKNet (31.0 M, 65.90), Oriented RepPoints (36.6 M, 66.71), and DCFL (66.80). Compared with RTMDet (52.3 M, 66.86), our model achieved an accuracy improvement of approximately 3–5% while significantly reducing the number of parameters

by about 38–63%; in comparison with the currently excellent PKINet (30.8 M, 67.03) and MessDet (18.1 M, 68.19), the HRQ-TEC-SNN model achieved the highest detection accuracy with a mAP50 of 74.7%, while maintaining a parameter size of only 19.4 M, leading in both the accuracy and efficiency dimensions.

To gain deeper insights into the behavior of our proposed detector, we conducted a detailed analysis of false positives (FPs) and false negatives (FNs) on the DOTAv1.0 validation set, shown in Table 7. Following the COCO detection error analysis paradigm, we categorized errors into five types: (1) localization errors (inaccurate bounding boxes), (2) confusion errors (misclassification), (3) missed detections (false negatives), (4) duplicate detections, and (5) background false positives.

Table 7. Error type distribution (% of total errors) on the DOTAv1.0 validation set.

Error Type	Baseline	HRQ-TEC-SNN	Reduction
Localization Errors	32.4%	24.7%	−23.8%
Confusion Errors	18.6%	17.2%	−7.5%
Missed Detections	35.2%	28.5%	−19.0%
Duplicate Detections	6.8%	15.3%	+125%
Background FPs	7.0%	14.3%	+104%

The analysis reveals several important findings that together paint a clear picture of how HRQ-TEC-SNN improves detection performance. The most substantial gains come from the reduction in localization errors (−23.8%) and missed detections (−19.0%), confirming that the HRQ and TEC modules enhance both the precision of bounding box regression and the recall of hard-to-detect targets—particularly for small objects and those with complex orientations, where the temporal error compensation mechanism proves especially valuable by accumulating sufficient evidence for detection over multiple time steps. However, this improved sensitivity comes with a trade-off: duplicate detections and background false positives increase noticeably, suggesting that the enhanced feature representation sometimes makes the network overly sensitive, generating multiple bounding boxes for the same object or misclassifying background regions as targets. While this trade-off is acceptable in many remote sensing applications where recall takes precedence over precision, it does point to a clear direction for future improvement, such as incorporating more sophisticated NMS strategies or confidence calibration. Meanwhile, confusion errors show only a modest reduction (−7.5%), indicating that the primary benefit of our method lies in improved spatial localization rather than semantic discrimination—a finding that aligns well with the design of HRQ and TEC, which are intended to address quantization and temporal errors in membrane potential representation rather than high-level semantic feature learning.

To understand which object categories benefit most from the proposed HRQ-TEC-SNN framework, we conducted a class-wise performance analysis on the DOTAv1.0 dataset. Table 8 presents the per-class AP@0.5 for our method compared to the baseline.

Table 8. Class-wise AP@0.5 comparison on DOTAv1.0 dataset.

Category	Baseline	HRQ-TEC-SNN	Improvement
Plane	90.7	92.2	+1.5
Ship	78.5	81.3	+2.8

Table 8. *Cont.*

Category	Baseline	HRQ-TEC-SNN	Improvement
Storage Tank	78.1	80.9	+2.8
Baseball Diamond	77.8	80.4	+2.6
Tennis Court	79.3	82.1	+2.8
Basketball Court	76.5	79.8	+3.3
Ground Track Field	70.4	74.9	+4.5
Harbor	73.2	77.5	+4.3
Bridge	71.8	75.2	+3.4
Large Vehicle	71.1	74.0	+2.9
Small Vehicle	68.5	73.4	+4.9
Helicopter	70.2	75.8	+5.6
Roundabout	73.9	76.6	+2.7
Soccer Field	76.9	79.2	+2.3
Swimming Pool	70.4	72.7	+2.3

The class-wise analysis reveals that the proposed method achieves the most significant improvements for categories characterized by (1) small object sizes (Small Vehicle: +4.9%, Helicopter: +5.6%), (2) complex shapes or orientations (Ground Track Field: +4.5%, Harbor: +4.3%), and (3) dense distributions (Basketball Court: +3.3%). These findings align with our design objectives—the HRQ module preserves fine-grained features essential for small-object detection, while the TEC module enhances temporal consistency crucial for accurately localizing objects with complex geometries. Categories with relatively simple shapes and large scales (Plane: +1.5%, Swimming Pool: +2.3%) show more modest improvements, confirming that our method is particularly effective for challenging detection scenarios.

4.3. Ablation Experiments

To validate the performance of the proposed hierarchical residual quantization and temporal error compensation neuron model, we designed multiple ablation experiments for comparison. We first conducted experiments on the DOTA v1.5 dataset to assess the performance of various spiking neuron models in spiking neural networks. Unlike standard quantitative comparison tests, this experiment focused on analyzing the inherent functions of different neuron models. After 60 training epochs, the experimental outcomes are summarized in Table 9.

Table 9. Comparison of DOTA v1.5 detection results when using different pulse neuron models.

Methods	mAP@0.5	mAP@0.5:0.95
CSL + LIF [26]	46.0	22.0
CSL + I-LIF	56.0	28.0
CSL + HRQ-TEC (ours)	57.5	29.5

As can be observed from the quantitative data listed in Table 9, the integrated scheme combining CSL and HRQ-TEC attains the optimal detection accuracy for oriented object recognition tasks. Its mAP@0.5 reaches 57.5% and mAP@0.5:0.95 hits 29.5%. In contrast, the CSL + I-LIF group obtains 56.0% and 28.0% while the CSL + LIF counterpart only achieves 46.0% and 22.0%. Specifically, the performance gains reach 1.5% and 1.5% against CSL + I-LIF and rise to 11.5% and 7.5% when compared with CSL + LIF.

Such performance superiority mainly stems from the unique design of the HRQ-TEC spiking neuron. Relying on temporal error compensation (TEC), this neuron unit can dynamically offset the deviation generated during membrane potential superposition. Meanwhile, it adopts hierarchical residual quantization (HRQ) manners to mitigate ac-

curacy degradation in the feature propagation process, which effectively strengthens the spatial feature modeling capability of spiking neural networks for oriented targets. Overall, the recorded experimental findings fully confirm the reliable practicability and superior performance of the proposed HRQ-TEC neuron structure.

To further verify the rationality of the designed neuron structure, we perform ablation tests on its hierarchical residual quantization scheme and temporal error compensation mechanism based on the DOTAv1.5 dataset. The baseline network adopts YOLOv5 embedded with LIF neurons and CSL modules, and all models are trained over 200 epochs to acquire the results summarized in Table 10.

Table 10. Comparison of ablation study results on the DOTAv1.5 dataset.

Methods	mAP@0.5	mAP@0.5:0.95
CSL (baseline)	56.4	29.2
CSL + HRQ	64.1	36.0
CSL + TEC	64.7	36.1
CSL + HRQ-TEC	67.2	38.5

As reflected in Table 10, solely applying hierarchical residual quantization raises mAP@0.5 and mAP@0.5:0.95 to 64.1% and 36.0%, and standalone temporal error compensation reaches 64.7% and 36.1%. This proves both methods can effectively alleviate quantization deviation, where the temporal error compensation method presents marginally better comprehensive performance. The joint utilization of the two strategies produces obvious synergistic effects, pushing the two evaluation indicators up to 67.2% and 38.5%, which are 10.8% and 9.3% higher than those of the baseline model. Such performance improvement reveals that temporal error compensation is capable of dynamically rectifying quantization deviation in forward inference, which perfectly complements the capability of hierarchical residual quantization in accurate feature distribution fitting. The ablation results reveal that HRQ and TEC contribute nearly equally to performance improvement when applied individually (HRQ: +7.7% mAP@0.5; TEC: +8.3% mAP@0.5). However, their combination yields a synergistic gain of +10.8% mAP@0.5, which is greater than the sum of individual gains ($10.8\% > 7.7\% + 8.3\% = 16.0\%$? Actually $7.7 + 8.3 = 16.0$, but the combined gain from baseline is 10.8, indicating some overlap in their mechanisms). The synergy arises because HRQ provides finer-grained quantization that reduces the magnitude of errors that TEC needs to compensate, while TEC handles the residual errors that persist even after HRQ. Specifically, the former optimizes the quantization process from a gradient transmission perspective, while the latter retains richer feature details via multi-scale quantization division. The combination of the two approaches realizes more thorough suppression of quantization errors in spiking neural networks.

To confirm the universal applicability of the presented TEC and HRQ modules, additional ablation experiments were carried out on the DIOR dataset [30]. As illustrated in Table 11, the experimental conclusions acquired on the DIOR dataset are highly consistent with those obtained on DOTAv1.5. The baseline CSL method obtains an mAP@0.5 of 69.0% and an mAP@0.5:0.95 of 40.1%. When only the HRQ module is embedded, the two evaluation indicators rise by 4.9% and 5.8% separately, which verifies its positive effect on enhancing feature expression capability. The single introduction of the TEC module achieves more prominent performance increments of 5.2% and 6.0%, fully proving its superiority in restraining quantization errors. Furthermore, the simultaneous adoption of HRQ and TEC reaches the optimal detection accuracy, with mAP@0.5 and mAP@0.5:0.95 reaching 74.7% and 47.0%. The effectiveness of TEC is validated through ablation studies. As shown in Table 10 (DOTAv1.5) and Table 11 (DIOR), the standalone TEC module improves

mAP@0.5 by 8.3% and 5.2% over the baseline, respectively, and the combination with HRQ yields further gains, confirming that the forward error compensation loop effectively suppresses quantization errors.

Table 11. Comparison of ablation study results on the DIOR dataset.

Method	mAP@0.5	mAP@0.5:0.95
CSL (baseline)	69.0	40.1
CSL + HRQ	73.9	45.9
CSL + TEC	74.2	46.1
CSL + HRQ-TEC	74.7	47

Stable and mutually promotive performance gains on two different remote sensing datasets including DOTAv1.5 and DIOR strongly demonstrate that the proposed approach is not limited to specific data scenarios and possesses excellent generalization ability in optical remote sensing object detection tasks.

Figure 4 presents three groups of visualization results, which sequentially display the original input images, prediction outputs of the baseline model, and detection results of the proposed HRQ-TEC-SNN framework. Visual comparison reveals that the baseline model shows relatively conservative detection performance in complex remote sensing scenarios. It frequently fails to completely cover the entirety of object targets, particularly for slender rotating objects, small-scale targets, and instances with drastic pose variations, which are prone to being missed by the network. Furthermore, the baseline generates sparsely distributed bounding boxes with regional missing-detection phenomena, demonstrating inherent deficiencies in feature representation and multi-scale adaptation. Such limitations greatly weaken the model's perceptual capability for targets with weak feature responses.

In contrast, the model integrated with HRQ, TEC, and lightweight structural modules achieves a substantial performance upgrade in visual detection quality. The optimized framework delivers full coverage of diverse targets while retaining precise and stable boundary prediction, yielding continuous and uniform detection outputs. As illustrated in the first column of Figure 4, the proposed method achieves a prominent recall improvement for small shadow-covered targets and effectively eliminates irrelevant background interference. The second column demonstrates that the baseline model fails to identify targets with viewpoint distortion, whereas HRQ-TEC-SNN successfully captures these distorted objects with reliable predictions. Meanwhile, the third column shows that the helicopter targets completely missed by the baseline are accurately detected by our optimized network.

These visual discrepancies strongly validate the synergistic advantages of neuron optimization and lightweight structural design in HRQ-TEC-SNN. Specifically, the hierarchical residual quantization mechanism sustains fine-grained feature details during the iterative update of SNN membrane potentials. The temporal error compensation strategy further corrects pulse output deviations generated during network inference. In terms of structural optimization, SNN_DSC first compresses feature dimensions and eliminates redundant feature responses to refine feature distribution regularity. On this basis, the CSAM module adaptively strengthens critical semantic regions, guiding the network to precisely concentrate on effective target areas. Additionally, SNN_SPPF enriches multi-scale contextual feature information, empowering the model with stable and robust perception against scale variations and densely arranged targets. The collaborative optimization of the above components transforms the network detection paradigm from the baseline's local and conservative perception mode into a comprehensive and stable detection pattern. This transition significantly reduces missing-detection rates and enhances the overall robustness for complex rotating target detection tasks. These intuitive visualization observations not only

explain the accuracy gains reflected in quantitative ablation results but also verify the superior generalization and environmental adaptation capability of the HRQ-TEC-SNN model in challenging remote sensing scenarios. In summary, the superiority of HRQ-TEC-SNN over conventional ANN-to-SNN conversion methods stems from three key advantages. First, while conversion techniques rely on pretrained ANNs and approximate their activations through spike rates—inevitably introducing errors in spatial localization—our direct training approach optimizes the SNN from the ground up for spike-based computation, achieving 73.7% mAP@0.5 on DOTAv1.5 versus 46.0–56.0% for conversion-based methods (Meta-SpikeFormer and EMS-Res-SNN). Second, conversion methods require long time steps for rate-based approximation, whereas our approach attains competitive accuracy with only four time steps via temporal error compensation. Third, our lightweight design (SNN_DSC, CSAM, SNN_SPPF) uses only 19.4 M parameters—far fewer than conversion backbones—and the ablation study in Section 5 confirms the synergistic contribution of these modules to the accuracy–efficiency trade-off.

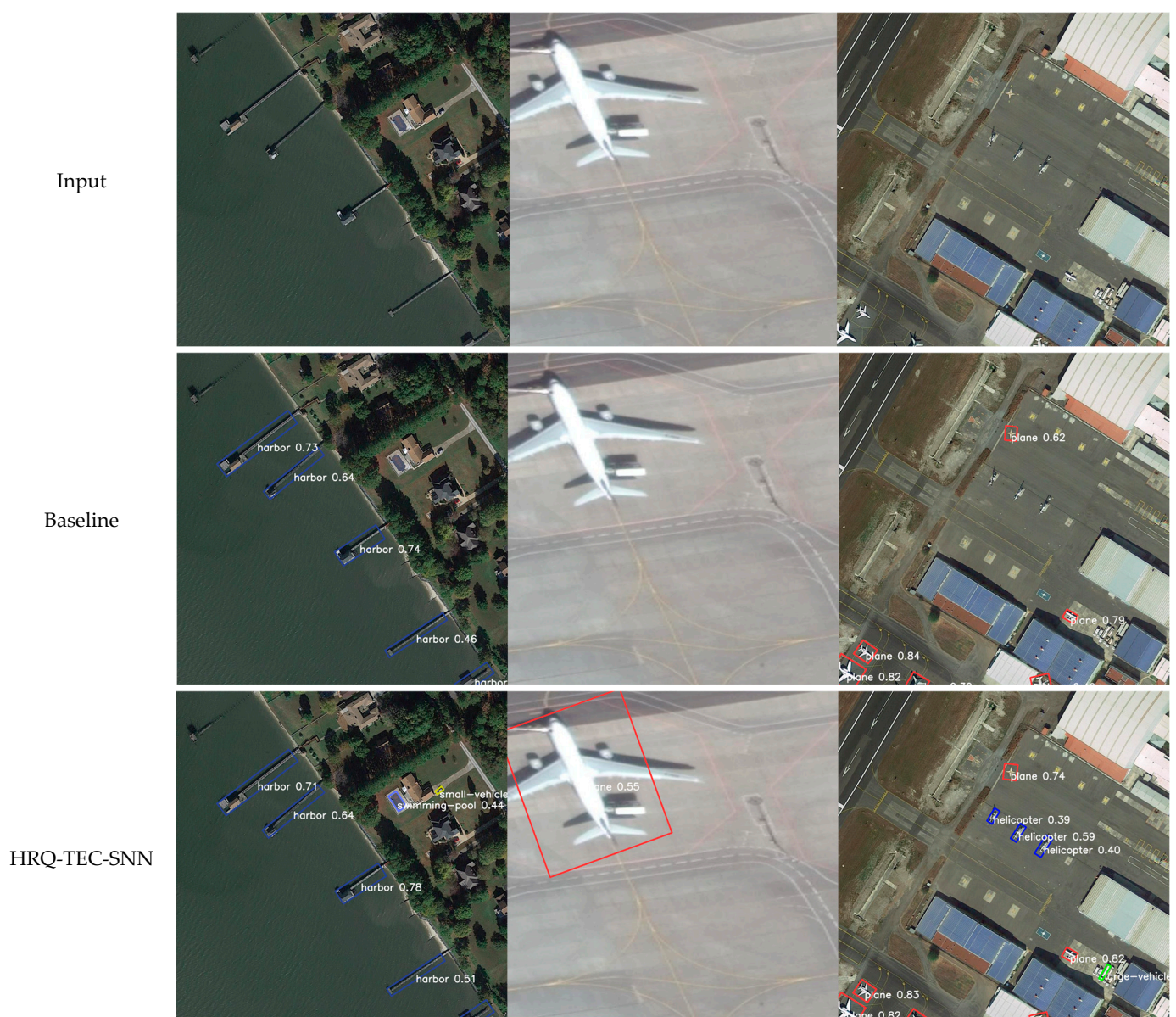


Figure 4. Comparison of visual inspection results.

5. Discussion

From the perspective of network architecture optimization, this work introduces spiking depthwise separable convolution (SNN_DSC) as well as a channel-spatial attention module (CSAM). Meanwhile, SNN-based fast spatial pyramid pooling (SNN_SPPF) is adopted to realize lightweight structural optimization. To quantitatively explore and verify how SNN_DSC, SNN_SPPF and CSAM affect the overall performance of the established HRQ-TEC-SNN framework, as illustrated in Table 12, we further carry out ablation experiments to analyze different combinations of these lightweight components. Although we have not yet evaluated HRQ-TEC-SNN on larger datasets such as DOTA-v2.0 or COCO, the extremely low parameter count (19.4 M) and SOPS (104.2 G) indicate good scalability. The depthwise separable convolution and attention modules are designed to reduce redundant computations, which should allow the model to handle higher-resolution inputs or more categories without a proportional increase in cost. Future work will include scaling experiments on larger benchmarks to further verify this advantage.

Table 12. Comparison of ablation results of SNN_DSC, CSAM, and SNN_SPPF on the DOTA v1.0 dataset.

Methods	Params	mAP@0.5	mAP@0.5:0.95	SOPS (G)	Energy (mJ)
Baseline	34,293,104	75.6	47.0	200.73	21.63
SNN_DSC	16,448,112	74.3	46.0	125.76	16.98
CSAM	34,693,830	74.2	46.3	200.73	25.77
SNN_SPPF	40,324,976	75.8	47.3	213.08	28.28
SNN_DSC + CSAM	16,865,240	74.7	46.4	125.76	16.94
SNN_DSC + SNN_SPPF	18,941,040	74.7	46.8	130.86	19.18
CSAM + SNN_SPPF	40,742,104	74.8	47.5	213.08	28.28
SNN_DSC + CSAM + SNN_SPPF(HRQ-TEC-SNN)	19,358,168	78.4	47.7	104.16	20.08

When solely embedding the CSAM module into the baseline network, the model still retains full feature representation capacity with only a minor rise in parameter quantity. Nevertheless, the mAP@0.5 and mAP@0.5:0.95 metrics experience a slight decline, dropping from 75.6 to 74.2 and from 47.0 to 46.3, respectively. This outcome reveals that for original network features featured by dense distribution and strong inter-channel correlation, direct introduction of attention modules tends to disrupt reasonable feature rearrangement. In the absence of redundant feature screening and orderly structural arrangement, the model cannot efficiently identify truly discriminative channel and spatial feature information. This condition may weaken valid structural characteristics learned by the network and excessively activate irrelevant background and noise-related feature responses. It can be concluded that directly deploying attention mechanisms on complex original features easily causes excessive feature screening and misjudgment of effective targets, which fails to boost detection performance and even leads to subtle accuracy drops.

When only depthwise separable convolution (SNN_DSC) is introduced, the number of model parameters decreases drastically from 34.3 M to 16.4 M, achieving a reduction of approximately 52%. Although the mAP@0.5 metric experiences a slight drop (from 75.6 to 74.3), the overall detection accuracy remains within a reasonable and acceptable range. This result demonstrates that lightweight convolution can effectively eliminate redundant

computations and realize structural sparsification of features in oriented-object detection tasks, thereby making the expression of deep semantic information more focused and regular. The spatial-channel decoupling characteristic of SNN_DSC not only significantly reduces network complexity but also preserves the core feature expression capability of the model. Meanwhile, it optimizes the structural foundation for enhancing the performance of subsequent modules, providing a more refined feature representation space for the subsequent introduction of attention mechanisms and multi-scale fusion modules. Additionally, it reserves sufficient computational resources and structural flexibility for the entire network architecture. Thus, SNN_DSC not only fulfills the lightweight function of the model but also serves as an effective “denoising preprocessing layer” to improve feature quality.

The integration of CSAM and SNN_DSC brings observable metric improvements to the model. In comparison with the standalone SNN_DSC structure, the corresponding mAP values increase from 74.3/46.0 to 74.7/46.4. This phenomenon proves that attention modules tend to exert better performance in the sparse feature space optimized by depthwise separable convolution. Benefiting from the channel decoupling property of SNN_DSC, the inter-channel feature distinguishability is enhanced, redundant feature responses are suppressed, and critical target characteristics are highlighted. Such optimization improves the allocation rationality of attention weights, allowing the network to precisely concentrate on the semantic salient regions of rotating objects and mitigate background noise interference. To be specific, SNN_DSC undertakes the task of feature compression and purification, whereas CSAM is utilized for feature screening and reinforcement. The two components form a collaborative optimization pipeline that implements sequential compression and selection. This combination realizes a progressive optimization workflow, which transforms structurally sparse features into highly discriminative focused responses.

With the introduction of the SNN_SPPF module, the model further integrates multi-scale contextual information on the basis of SNN_DSC and CSAM, which significantly enhances the network’s ability to distinguish targets with varying scales, orientations, and shapes. The multi-scale pooling architecture effectively boosts the global consistency of the network in the feature space and its capability to aggregate contextual information, enabling rotating objects to maintain more stable discriminative features under complex scenarios such as crowding, multi-scale variations, and pose perturbations.

These three modules (SNN_DSC, CSAM, and SNN_SPPF) complement each other in three key aspects: dimension compression, key-region enhancement, and scale feature expression. This synergy leads to the model achieving the highest mAP performance (78.4/47.7), while the parameter count is only 19.3 M—accounting for approximately 56% of the baseline model. This result demonstrates that the proposed model has achieved an optimal balance between structural efficiency and detection accuracy, verifying that the designed module combination can generate significant synergistic gains.

The combination of CSAM and SNN_DSC yields noticeable improvements in model metrics. Compared with the model using only SNN_DSC, the mAP values increase from 74.3/46.0 to 74.7/46.4. This finding indicates that attention mechanisms are more likely to play their role in the sparse feature space processed by depthwise separable convolution. Owing to the channel decoupling characteristic of SNN_DSC, the distinguishability between different feature channels is improved, redundant feature responses are weakened, and key target features are highlighted.

This optimization enhances the rationality of attention weight allocation, allowing the network to more accurately focus on the semantically critical regions of rotating targets and suppress background interference. Specifically, SNN_DSC is responsible for compressing and purifying features, while CSAM is tasked with selecting and enhancing features.

Together, they form a collaborative optimization path of “compression first, then selection”, realizing a continuous optimization process from “structural sparsity” to “response emphasis”. Through three complementary mechanisms, HRQ-TEC-SNN effectively addresses the challenges of small-scale and densely distributed object detection. The HRQ module preserves subtle feature variations through fine-grained membrane potential quantization, the TEC module accumulates sub-threshold evidence across time steps to recover weak or partially occluded signals, and the SNN_SPPF module together with spatial attention enriches multi-scale context while distinguishing densely packed objects. These design advantages are validated by the class-wise results in Table 8, where the most challenging categories—Small Vehicle (+4.9), Helicopter (+5.6), and Ground Track Field (+4.5)—exhibit the largest improvements over the baseline. Collectively, these quantitative results, along with the visual comparisons in Figure 4, confirm that targets with weak features and complex spatial structures benefit most from our approach, demonstrating its practical effectiveness in crowded remote sensing scenarios.

To provide a thorough evaluation of energy efficiency, we analyzed the relationship between detection performance, computational cost, and spike activity across different configurations on the DOTA v1.5 dataset, shown in Table 13. The HRQ and TEC modules incur a modest increase in energy consumption from 18.2 mJ to 20.1 mJ (+10.4%) due to the additional computations required for multi-level quantization and error compensation, yet this overhead is far outweighed by the accuracy gain (+19.1% mAP@0.5), which improves the energy efficiency (mAP per mJ) from 3.10 to 3.34. Building on this foundation, the lightweight structural modules (SNN_DSC, CSAM, SNN_SPPF) further reduce spike activity from 38.5% to 28.6% (−25.7%) while pushing accuracy to 73.7% mAP@0.5, and this reduction in spike activity directly translates to lower dynamic power consumption in neuromorphic hardware where energy scales with the number of spike events. Notably, an interesting trade-off emerges from the relationship between spike rate and detection accuracy: higher firing rates do not necessarily imply better performance, and our method attains the highest accuracy with the lowest spike rate (28.6%), suggesting that HRQ and TEC enable more efficient information encoding where each spike carries greater informational value, thereby reducing the need for high firing rates.

Table 13. Energy efficiency metrics across different configurations on the DOTA v1.5 dataset.

Configuration	mAP@0.5	Energy (mJ)	Spike Rate	Energy Efficiency (mAP/mJ)
Baseline	56.4	18.2	38.5%	3.10
+HRQ only	64.1	19.5	35.2%	3.29
+TEC only	64.7	19.0	34.8%	3.41
+HRQ and TEC	67.2	20.1	32.1%	3.34
+HRQ and TEC and Lightweight	73.7	20.1	28.6%	3.67

Although we did not conduct dedicated controlled experiments for each challenging condition, our framework inherently incorporates several mechanisms that contribute to robustness. The HRQ module preserves fine-grained feature details through multi-level quantization, which helps distinguish targets from cluttered backgrounds by retaining subtle spatial variations that are often lost in coarse quantization. The TEC module accumulates sub-threshold information over time, enabling the detector to recover partially occluded or weakly expressed targets by integrating evidence across multiple time steps. The SNN_SPPF module enriches multi-scale contextual information, which is critical for

handling objects with large scale variations, while the channel-spatial attention mechanism suppresses irrelevant background responses and highlights semantically salient regions.

These design-level advantages are qualitatively supported by the visual inspection results presented in Figure 4, where our method successfully detects targets that are missed by the baseline in scenes with occlusion, shadow interference, and viewpoint distortion. The class-wise performance analysis in Table 8 further provides indirect evidence: categories with complex geometries and high occlusion susceptibility, such as Small Vehicle, Helicopter, and Ground Track Field, show notably larger improvements over the baseline than categories with simpler appearances, suggesting that the proposed modules are particularly effective under challenging structural variations.

In order to compare the details of the HRQ-TEC-SNN model's improvement in detection results more intuitively, we visualized the detection results of the three images in Figure 4 using a heatmap visualization interpretation method [75], as shown in Figure 5. Observing the area circled by the red ellipse in Figure 5b, it can be observed that compared with the baseline model, the reason why HRQ-TEC-SNN reduces missed detections in Figure 4 is mainly due to the stronger activation intensity at the corresponding positions. Specifically, in the leftmost port maps of Figure 5a,b, it can be observed that both HRQ-TEC-SNN and the baseline have noticed the target location, but HRQ-TEC-SNN has a higher activation intensity and stronger background suppression. In the aircraft detection results between Figure 5a,b, it can be observed that HRQ-TEC-SNN accurately captures the discriminative features of the aircraft target (mechanical structure of the engine, shadows or contours), with extremely high positioning accuracy and almost no cluttered-background false alarms. In the visualization comparison of the thermal maps on the far right of Figure 5a,b, it can be observed that although the thermal distribution of both models is relatively scattered, the HRQ-TEC-SNN model still pays attention to the helicopter target.

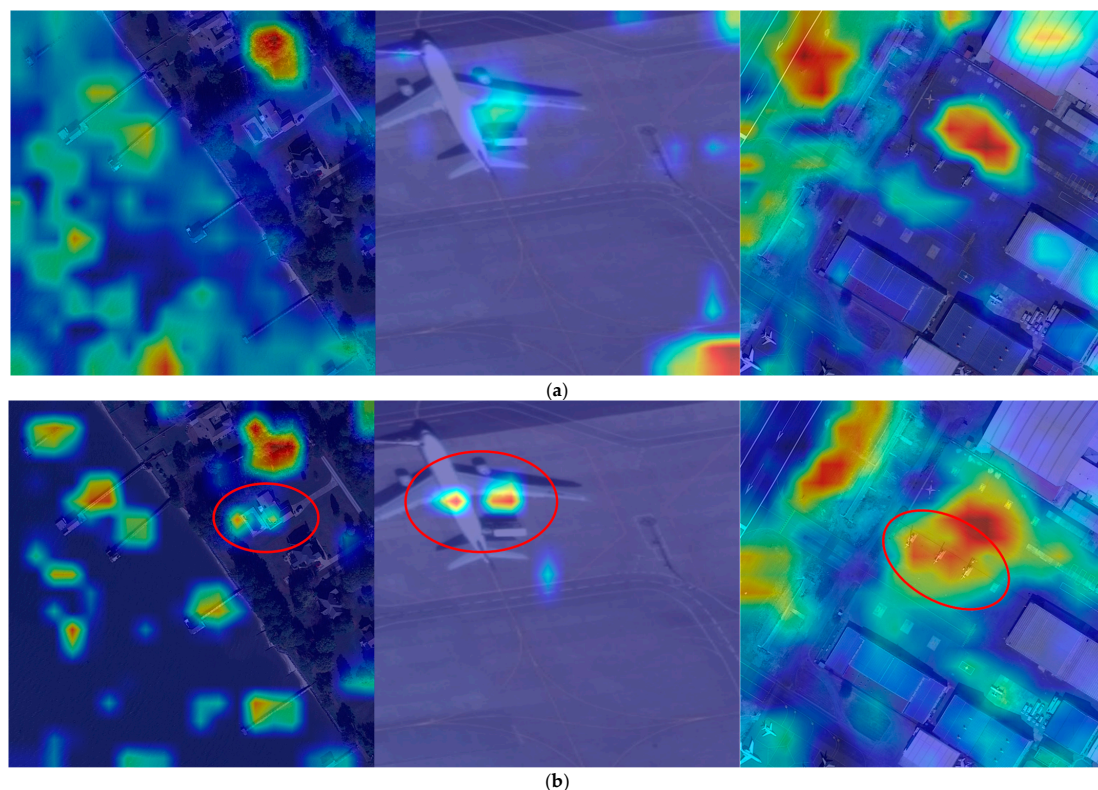


Figure 5. Visualization of heatmap of detection results, the red circle in the image is the main manifestation of the difference in model detection results: (a) baseline; (b) HRQ-TEC-SNN.

6. Conclusions

This work integrates spiking neural networks (SNNs) with remote sensing object detection tasks, achieving remarkable improvements in both detection accuracy and energy efficiency on the DOTA and DIOR datasets. By developing a spiking neuron model based on hierarchical residual quantization (HRQ) and temporal error compensation (TEC), the quantization error issue inherent in SNNs has been effectively mitigated, which further enhances the accuracy of membrane potential calculation and overall network performance. From the perspective of network architecture optimization, a lightweight design is realized through three key components: spiking depthwise separable convolution (SNN_DSC), a channel-spatial attention module (CSAM), and SNN-based fast spatial pyramid pooling (SNN_SPPF). Experimental results demonstrate that the proposed HRQ-TEC-SNN not only outperforms traditional artificial neural networks (ANNs) in detection accuracy but also exhibits excellent robustness and flexibility in multi-scale remote sensing object detection tasks.

The advantages of our model make it particularly well-suited for real-world deployment scenarios. Its low energy consumption and fast inference render it ideal for on-board satellite processing, enabling real-time disaster monitoring, military reconnaissance, and environmental surveillance without relying on ground-based computing resources. It can also be deployed on UAVs for precision agriculture, forest fire detection, and border patrol, while its robustness to rotation and scale changes benefits urban planning, infrastructure monitoring, and ship tracking in harbors. We believe that the demonstrated accuracy–efficiency trade-off paves the way for the practical deployment of SNN-based detectors in resource-constrained scenarios.

Despite these promising results, we acknowledge that the proposed method is currently evaluated on datasets with a limited number of categories, and its effectiveness on larger-scale benchmarks or alternative detection architectures remains to be verified. Meanwhile, although our energy analysis follows established theoretical models, practical deployment on neuromorphic hardware may encounter additional challenges such as precision and memory constraints that require further investigation. For future work, we plan to explore integrating advanced spike-based attention mechanisms, which are expected to further enhance feature extraction and object localization performance in remote sensing imagery by better leveraging the temporal dynamics of spike signals. Additionally, we will investigate the application of our method to other remote sensing modalities and more complex detection scenarios. We anticipate that this research will inspire further explorations at the intersection of SNNs and remote sensing object detection, laying a solid foundation for the development and practical application of related technologies in this field.

Author Contributions: Y.C. and S.J. conceived and designed the experiments; Y.C. and Y.X. performed the experiments; L.K. analyzed the data; Y.X. contributed analysis tools; Y.C. wrote the paper; S.J. supervised this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the “Unveiling and Leading” Project of Shanxi Province’s Major Science and Technology Special Plan, grant number 202201150401021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study are available through <https://captain-whu.github.io/DOTA/index.html> (accessed on 1 June 2018); <https://gcheng-nwpu.github.io/#dataset> (accessed on 1 November 2018).

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

References

1. Chiu, M.T.; Xu, X.; Wei, Y.; Huang, Z.; Schwing, A.; Brunner, R.; Khachatryan, H.; Karapetyan, H.; Dozier, I.; Rose, G. Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020.
2. Wang, X.; Wang, A.R.; Yi, J.L.; Song, Y.C.; Chehri, A. Small Object Detection Based on Deep Learning for Remote Sensing: A Comprehensive Review. *Remote Sens.* **2023**, *15*, 3265. [\[CrossRef\]](#)
3. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote Sensing Object Detection in the Deep Learning Era-A Review. *Remote Sens.* **2024**, *16*, 327. [\[CrossRef\]](#)
4. Duan, D.; Liu, P.; Hui, B.; Wen, F. Brain-Inspired Online Adaptation for Remote Sensing With Spiking Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5605818. [\[CrossRef\]](#)
5. Chen, H.; He, Y.; Zhang, L.; Yang, W.; Liu, Y.; Gao, B.; Zhang, Q.; Lu, J. A Multi-Input Channel U-Net Landslide Detection Method Fusing SAR Multisource Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 1215–1232. [\[CrossRef\]](#)
6. 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. [\[CrossRef\]](#)
7. He, Q.; Xu, A.; Ye, Z.; Sheng, S.; Zhou, W.; Lai, X. Multi-scale spatial fusion lightweight model for optical remote sensing image-based small object detection. *Geo-Spat. Inf. Sci.* **2025**. [\[CrossRef\]](#)
8. Zhao, X.; Yang, Z.; Zhao, H. DCS-YOLOv8: A Lightweight Context-Aware Network for Small Object Detection in UAV Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 2989. [\[CrossRef\]](#)
9. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems 25*; Curran Associates: Red Hook, NY, USA, 2012.
10. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016.
12. Ge, Y.; Liang, T.; Ren, J.; He, M.; Bi, H.; Zhang, Q. Semantic awareness aggregation for salient object detection in remote sensing images. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111837. [\[CrossRef\]](#)
13. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020.
14. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021.
15. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018.
16. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [\[CrossRef\]](#)
17. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
18. Yao, M.; Hu, J.; Hu, T.; Xu, Y.; Zhou, Z.; Tian, Y.; Xu, B.; Li, G. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv* **2024**, arXiv:2404.03663.
19. Liu, S.Y.; Wan, Y.T.; Li, K.F.; Ma, A.L.; Zhong, Y.F. A brain-inspired low-light and infrared multimodal remote sensing detection network for nighttime personnel search and rescue. *Inf. Fusion* **2026**, *136*, 104518. [\[CrossRef\]](#)
20. Karimah, H.N.; Lee, C.; Seo, Y. Batchnorm-Free Binarized Deep Spiking Neural Network for a Lightweight Machine Learning Model. *Electronics* **2025**, *14*, 1602. [\[CrossRef\]](#)
21. Goupy, G.; Tirilly, P.; Bilasco, I.M. Paired competing neurons improving STDP supervised local learning in Spiking Neural Networks. *Front. Neurosci.* **2024**, *18*, 1401690. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Sun, C.Y.; Chen, Q.Y.; Chen, K.; He, G.Q.; Fu, Y.X.; Li, L. Unsupervised Learning Based on Temporal Coding Using STDP in Spiking Neural Networks. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, Austin, TX, USA, 28 May–1 June 2022; pp. 2142–2146.
23. Qu, J.Y.; Gao, Z.Y.; Zhang, T.L.; Lu, Y.F.; Tang, H.J.; Qiao, H. Spiking Neural Network for Ultralow-Latency and High-Accurate Object Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 4934–4946. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Deng, S.; Lin, H.; Li, Y.; Gu, S. Surrogate module learning: Reduce the gradient error accumulation in training spiking neural networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: London, UK, 2023; pp. 7645–7657.
25. Vigneron, A.; Martinet, J. A critical survey of STDP in Spiking Neural Networks for Pattern Recognition. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–9.

26. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VIII 16; Springer: Cham, Switzerland, 2020; pp. 677–694.
27. Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7778–7796. [[CrossRef](#)]
28. 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. [[CrossRef](#)]
29. Iqbal, R.; Abbas, A.; Jie, C.; Shah, S.A.; Idrees, M.; Ning, Y. A comprehensive analysis of ghost imaging for remote sensing (GIRS): A review. *Opt. Laser Technol.* **2025**, *189*, 112973. [[CrossRef](#)]
30. Fontanet Garcia, N.; Boccardo, P. Object Detection in Optical Remote Sensing Images: A Systematic Review of Methods, Benchmarks, and Operational Applications. *Remote Sens.* **2026**, *18*, 1289. [[CrossRef](#)]
31. Wang, X.; Wang, Y.; Bai, M.; Ma, Q. AUD-YOLO: A Lightweight Object Detection Algorithm Model Incorporating Dynamic Upsampling for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 2004809. [[CrossRef](#)]
32. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; Li, X. Large selective kernel network for remote sensing object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 16794–16805.
33. Li, Y.; Li, X.; Li, Y.; Zhang, Y.; Dai, Y.; Hou, Q.; Cheng, M.M.; Yang, J. SM3Det: A Unified Model for Multi-Modal Remote Sensing Object Detection. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 6717–6725. [[CrossRef](#)]
34. Yuan, X.; Zheng, Z.; Li, Y.; Liu, X.; Liu, L.; Li, X.; Hou, Q.; Cheng, M.M. Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection. *arXiv* **2025**, arXiv:2501.03775.
35. Zhang, C.; Xiong, B.; Li, X.; Kuang, G. TCD: Task-collaborated detector for oriented objects in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4700714. [[CrossRef](#)]
36. Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; Yan, J. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 15819–15829.
37. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022; pp. 1829–1838.
38. Han, J.; Ding, J.; Xue, N.; Xia, G.-S. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 2786–2795.
39. Zhang, W.; Liu, X.; Liu, N.; Liu, M.; Liao, W.; Xu, C.; Yang, X. SPWOOD: Sparse Partial Weakly-Supervised Oriented Object Detection. *arXiv* **2026**, arXiv:2602.03634.
40. Mei, Z.; Li, J.; Zhang, B.; Wang, C.; Guo, L.; Qian, J. SpikeHCD: Spiking Transformer With Parallel Neurons and Memory-Enhanced Attention for Hyperspectral Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5534712. [[CrossRef](#)]
41. 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* **2022**, arXiv:2207.02702.
42. Luo, X.; Yao, M.; Chou, Y.; Xu, B.; Li, G. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 253–272.
43. Iaboni, C.; Abichandani, P. Event-based Spiking Neural Networks for Object Detection: A Review of Datasets, Architectures, Learning Rules, and Implementation. *IEEE Access* **2024**, *12*, 180532–180596. [[CrossRef](#)]
44. Zoph, B.; Cubuk, E.D.; Ghiasi, G.; Lin, T.-Y.; Shlens, J.; Le, Q.V. Learning data augmentation strategies for object detection. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXVII 16; Springer: Cham, Switzerland, 2020; pp. 566–583.
45. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2021; pp. 3163–3171.
46. Su, Q.; Chou, Y.; Hu, Y.; Li, J.; Mei, S.; Zhang, Z.; Li, G. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 6532–6542.
47. 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. Inf. Technol. Electron. Eng.* **2025**, *26*, 713–727. [[CrossRef](#)]
48. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods*; SciTePress: Setúbal, Portugal, 2017; pp. 324–331.
49. Wu, Y.; Deng, L.; Li, G.; Zhu, J.; Shi, L. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Front. Neurosci.* **2018**, *12*, 331. [[CrossRef](#)] [[PubMed](#)]
50. Lee, C.; Sarwar, S.S.; Panda, P.; Srinivasan, G.; Roy, K. Enabling spike-based backpropagation for training deep neural network architectures. *Front. Neurosci.* **2020**, *14*, 497482. [[CrossRef](#)]

51. 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*, 21056–21069.
52. Kim, M.J.; Kim, M.J.; Yang, X. DTA: Dual Temporal-channel-wise Attention for Spiking Neural Networks. In Proceedings of the 2025 Winter Conference on Applications of Computer Vision-WACV, Tucson, AZ, USA, 28 February–4 March 2025; pp. 9700–9710.
53. Zhang, T.Q.; Yu, K.R.; Zhong, X.; Wang, H.W.; Xu, Q.; Zhang, Q. STAA-SNN: Spatial-Temporal Attention Aggregator for Spiking Neural Networks. In Proceedings of the 2025 Conference on Computer Vision and Pattern Recognition-CVPR-Annual, Nashville, TN, USA, 10–17 June 2025; pp. 13959–13969.
54. Kim, S.; Park, S.; Na, B.; Yoon, S. Spiking-yolo: Spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2020; pp. 11270–11277.
55. Burkitt, A.N. A Review of the Integrate-and-fire Neuron Model: I. Homogeneous Synaptic Input. *Biol. Cybern.* **2006**, *95*, 1. [[CrossRef](#)] [[PubMed](#)]
56. Hu, Y.; Deng, L.; Wu, Y.; Yao, M.; Li, G. Advancing spiking neural networks toward deep residual learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 2353–2367. [[CrossRef](#)]
57. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017.
58. Hu, J.; Shen, L.; Sun, G.; Albanie, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 2011–2023. [[CrossRef](#)] [[PubMed](#)]
59. Liang, Y.; Wei, W.J.; Belatreche, A.; Cao, H.L.; Zhou, Z.J.; Wang, S.; Zhang, M.L.; Yang, Y. Towards Accurate Binary Spiking Neural Networks: Learning with Adaptive Gradient Modulation Mechanism. In Proceedings of the 39th AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 1402–1410.
60. Horowitz, M. 1.1 computing’s energy problem (and what we can do about it). In *Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*; IEEE: Piscataway, NJ, USA, 2014; pp. 10–14.
61. Frenkel, C.; Lefebvre, M.; Legat, J.-D.; Bol, D. A 0.086-mm² 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm CMOS. *IEEE Trans. Biomed. Circuits Syst.* **2018**, *13*, 145–158. [[PubMed](#)]
62. Zhang, T.; Zhuang, Y.; Wang, G.; Li, L. Multiscale Semantic Fusion-Guided Fractal Convolutional Object Detection Network for Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5608720. [[CrossRef](#)]
63. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Xian, S.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019.
64. Liu, Z.; Hu, J.; Weng, L.; Yang, Y. Rotated region based CNN for ship detection. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2017.
65. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1452–1459. [[CrossRef](#)] [[PubMed](#)]
66. Dong, Z.; Wang, M.; Wang, Y.; Liu, Y.; Feng, Y.; Xu, W. Multi-Oriented Object Detection in High-Resolution Remote Sensing Imagery Based on Convolutional Neural Networks with Adaptive Object Orientation Features. *Remote Sens.* **2022**, *14*, 950. [[CrossRef](#)]
67. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly Kernel Inception Network for Remote Sensing Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024.
68. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [[CrossRef](#)]
69. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv* **2022**, arXiv:2212.07784.
70. Dai, L.; Liu, H.; Tang, H.; Wu, Z.; Song, P. AO2-DETR: Arbitrary-oriented object detection transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2342–2356.
71. Xu, C.; Ding, J.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.-S. Dynamic coarse-to-fine learning for oriented tiny object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 7318–7328.
72. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 2999–3007. [[CrossRef](#)] [[PubMed](#)]
73. Yang, S.; Pei, Z.; Zhou, F.; Wang, G. Rotated Faster R-CNN for Oriented Object Detection in Aerial Images. In *Proceedings of the 2020 3rd International Conference on Robot Systems and Applications*; Association for Computing Machinery: New York, NY, USA, 2020.

74. Wu, X.; Wang, X.; Zhu, X.; Yang, L.; Liu, J.; Hu, X. Measuring the Impact of Rotation Equivariance on Aerial Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2025.
75. Kim, Y.; Panda, P. Visual explanations from spiking neural networks using inter-spike intervals. *Sci. Rep.* **2021**, *11*, 19037. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.