SAGA-YOLO: A High-Accuracy Detector for SAR Aircraft in Complex Environments

Qing GUO[1,2], Hongdong ZHAO[1,2,], Wenjing WANG[1,2], Wenkai ZHANG[1,2]*

[1] School of Electronic Information Engineering, Hebei University of Technology, 300401 Tianjin, China
[2] Innovation and Research Institute of Hebei University of Technology in Shijiazhuang, 050299 Shijiazhuang, China

guoq5916@163.com, zhaohd@hebut.edu.cn, 481305864@qq.com, zwk999vip@163.com

Submitted April 29, 2026 / Accepted August 17, 2026 / Online first September 18, 2026

Abstract. *Synthetic Aperture Radar (SAR), utilizing its ability to actively transmit microwave signals, is widely used for military aircraft target reconnaissance and aviation safety monitoring. However, the complex background of ground-based aircraft targets, combined with the coherent imaging characteristics of SAR, results in speckle noise, which increases the difficulty of detection. To address this, we propose a high-accuracy SAR aircraft detection model, SAR-Adaptive Gated-Attention You Only Look Once (SAGA-YOLO). It includes the following three improvements: Firstly, the model constructs a robust gated convolutional backbone network that adaptively filters out noise and irrelevant clutter by dynamically adjusting feature channels, thereby extracting more robust semantic features of aircraft. Secondly, an attention module, C2-CAS, is introduced at the end of the backbone network to refocus features on the most distinctive scattering regions of the aircraft, enhancing its key structural features. Finally, a feature fusion network, Slim-neck, is introduced at the neck to improve the model's ability to extract and fuse features of targets at different scales. Experimental results on the SAR-AIRcraft-1.0 and SADD aircraft datasets demonstrate that SAGA-YOLO achieves superior performance; compared to the baseline YOLO11, SAGA-YOLO improves mAP50 and mAP50-95 by 3.0% and 5.8% respectively on SAR-AIRcraft-1.0, and by 0.6% and 18.3% on SADD.*

Keywords

SAR image, YOLO11, aircraft detection, gated backbone

1. Introduction

With the advancement of object detection technology, it has demonstrated excellent performance on natural images. However, in the case of SAR images, their coherent imaging characteristics give rise to speckle noise [1], and the distribution of scattering points from aircraft targets is discrete, making them susceptible to strong scattering interference from complex airport backgrounds. These factors all increase the difficulty of distinguishing aircraft targets from the background, posing significant challenges for target recognition tasks [2]. Consequently, the development of an efficient and robust SAR image target detection model is crucial for the accurate detection and recognition of specific targets such as airports and aircraft.

Currently, the academic community has conducted preliminary explorations into the application of deep learning methods for aircraft target detection in SAR imagery. For instance, Ke et al. [3] proposed the Scatter Feature-Guided Network (SFG-Net), which integrates feature extraction, global feature fusion and label assignment with the core scatter distribution of the target, thereby mitigating issues of blurred aircraft details and noise interference. Liu et al. [4] proposed the Scatter-Enhanced and Feature-Fusion Network (SEFF-Net), which employs a Scatter Information Extraction and Enhancement Module (SIEEM) to highlight aircraft target scatter points, coupled with a Spatio-Depth Coordinate Attention Module (SD-CAM) to focus on the aircraft target's location. Wang et al. [5] proposed the lightweight MSADNet, which innovatively incorporates a Multi-stage Asymmetric Aggregation Network (MAAN) module. By combining multi-level attention mechanisms with a Coordinate Attention (CA) module, it enhances both the accuracy and efficiency of aircraft recognition. Wang et al. [6] proposed an anchor-free SAR aircraft detection method based on density map prediction. By encoding aircraft labels as two-dimensional Gaussian density maps, this method effectively mitigates the issue of label overlap in dense target scenarios, enabling the detection of densely clustered aircraft targets.

These methods have improved the detection performance of aircraft targets in SAR images to some extent; however, they still suffer from shortcomings in terms of detection accuracy and speed, and face significant challenges in effectively minimizing information loss and adapting to complex background interference.

To address these issues, we propose SAGA-YOLO, a high-precision adaptive model specifically designed for the detection of aircraft targets at airports using SAR imagery. The main contributions are as follows:

1. We have constructed a novel hierarchical gated convolutional backbone network. By introducing a dynamic

DOI: 10.13164/re.2026.0442

spatial gating mechanism and deep separable convolutions with large kernels, we have enhanced the model's robustness in extracting aircraft features within complex airport environments at the architectural level.

2. We have innovatively designed the C2-CAS collaborative attention mechanism to act as a filter for deep-level features. Through a channel attention scaling mechanism, it accurately captures and focuses on strong scattering centers specific to aircraft targets, such as wings, thereby effectively distinguishing aircraft targets from metallic background clutter.

3. We have introduced an efficient multi-scale feature fusion architecture based on Slim-neck. Without causing parameter explosion, this effectively fuses deep semantic information from the backbone with shallow spatial details, thereby improving the recognition rate of aircraft at different scales.

4. To address the accuracy of aircraft recognition in the specific context of SAR imagery, we have constructed a comprehensive high-precision SAGA-YOLO model. Its recognition capabilities have been validated on the SAR-AIRcraft-1.0 and SADD datasets.

2. Related Works

2.1 Traditional SAR Image Object Detection Methods

Early SAR target detection methods were predominantly based on Constant False Alarm Rate (CFAR) [7], the core principle of which assumes that background clutter follows a Gaussian distribution. By estimating the statistical parameters of background clutter within a sliding window, an adaptive threshold is dynamically calculated to maintain a constant false alarm rate. To adapt to different clutter environments, researchers have subsequently proposed various variants, such as Ordered Statistical CFAR (OS-CFAR) [8] and two-parameter CFAR [9], among other variants, with the aim of maintaining a stable false alarm rate in both uniform and multi-target interference environments. However, this approach relies heavily on the statistical characteristics of the radar image and manual design; it suffers from poor real-time performance, and the expertise of the statistician can also influence the detection results.

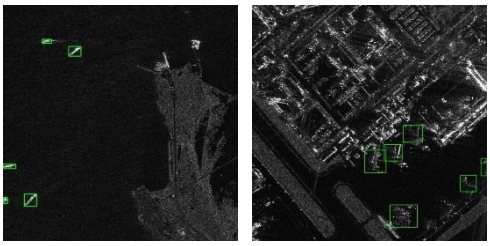
In addition to detection methods based on statistical distributions, detection schemes based on traditional machine learning and feature engineering also occupy an important position in the field of SAR target recognition. Such methods typically begin by identifying candidate target regions via CFAR or saliency detection, followed by the extraction of texture, edge and geometric structural features using techniques such as morphological filtering, wavelet transform, LBP, HOG or GLCM. Finally, classifiers such as Support Vector Machines (SVM) [10] and AdaBoost [11] are employed to distinguish targets from clutter, thereby completing the target detection of the candidate regions.

However, traditional methods have clear limitations when applied to high-resolution, complex-scene SAR aircraft detection tasks. In complex airport environments, strong scattering clutter from structures and transport vehicles leads to an increase in CFAR false alarms, resulting in extremely high false positive rates. Furthermore, the manually designed features in CFAR methods rely heavily on expert domain knowledge and struggle to cope with the significant variations in scale and discrete scattering patterns exhibited by aircraft targets under different radar incidence angles and resolutions.

2.2 A Deep Learning-Based SAR Target Detection Method

Currently, deep learning-based detection methods are primarily categorized into two-stage and single-stage approaches. Two-stage algorithms, exemplified by Faster R-CNN [12] and Cascade R-CNN [13], utilize a Region Proposal Network (RPN) to generate high-quality candidate bounding boxes, followed by precise classification and location regression. Although these methods have achieved high detection accuracy on SAR imagery, their computationally intensive logic and large number of parameters result in slow inference speeds, making it difficult to meet the real-time requirements of intelligent monitoring.

In contrast, single-stage detectors, represented by the YOLO [14] series and SSD [15], transform object detection into a direct regression problem, achieving an excellent balance between accuracy and inference speed, and have rapidly become the mainstream architecture in the field of SAR object detection. Many researchers have adapted the YOLO model to the characteristics of SAR imagery; Yang et al. [16] utilized the YOLOv4 object detection algorithm and publicly available Sentinel-1 SAR data to develop a deep learning-based oil spill detector for the eastern Mediterranean. Guo et al. [17] proposed the DS-YOLO model, which effectively improves the detection accuracy of small, densely clustered ships in SAR images by introducing SPD-Conv and CSP-PPA modules and optimizing the AWNWD bounding box loss function. Mo et al. [18] proposed LEAD-YOLO, a lightweight SAR vessel detection model based on YOLOv5. Through optimizations to the backbone and attention modules, it significantly reduces model complexity and improves detection speed whilst maintaining detection accuracy superior to most methods. Lejun et al. [19] proposed DGSP-YOLO, an enhancement of YOLOv8n, and thoroughly validated its high accuracy in SAR vessel recognition under complex sea conditions across three mainstream SAR datasets; Rocha et al. [20] implemented targeted improvements to YOLOv12, not only enhancing vessel recognition capabilities but also achieving notable results on aircraft and other multi-class SAR datasets. The AMDC-YOLO proposed by Wang L et al. [21] effectively enhances feature extraction capabilities for low-resolution SAR images by integrating BiFormer modules with DynamicConv modules, whilst also improving detection accuracy for small targets. Bayraktar et al. [22] added an ECA module following the feature fusion module in the neck of YOLO11 to enhance



(a) (b)

Fig. 1. SAR images of vessels and aircraft: (a) vessels; (b) aircrafts.

the detection performance of ground-based aircraft in SAR images.

Based on the above analysis, it can be seen that although YOLO-based improvements have achieved high accuracy in SAR target detection and made some progress in model lightweighting, most of these improvements currently focus on SAR vessel targets. As shown in Fig. 1. aircraft targets are typically located in more complex airport environments and exhibit more complex scattering patterns, thereby increasing the difficulty of target identification. In such complex scenarios, it is crucial to effectively distinguish between aircraft scattering patterns and background clutter in order to correctly identify aircraft targets.

3. Methodology

Given the coherent speckle noise generated by SAR's unique imaging mechanism, as well as the complex corner scattering characteristics of aircraft targets and interference from ground clutter at the airport, certain requirements are placed on the accuracy of the detection model. In addition, compared to two-stage detectors such as Faster R-CNN, single-stage detectors offer faster inference speeds, which is crucial for real-time SAR image processing in practical deployment scenarios. Taking these factors into account, we selected the classic YOLO11n method in object detection as the baseline model. As an improved version of YOLOv8n,

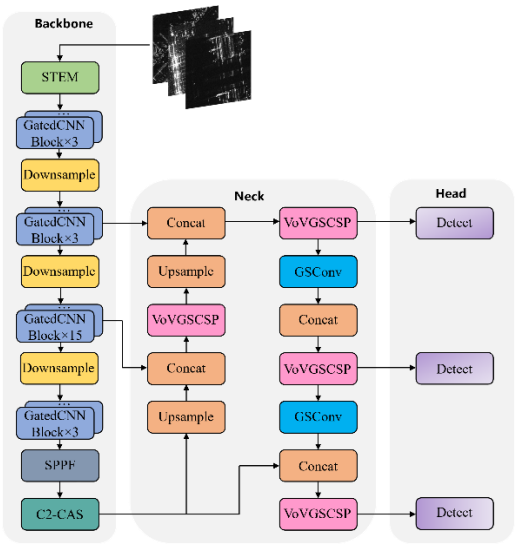


Fig. 2. SAGA-YOLO network structure.

YOLO11n offers higher detection accuracy with fewer parameters, and the introduction of the C2PSA module in the backbone enables better feature extraction for targets in SAR images, which are characterized by a large amount of coherent speckle noise and strong scattering interference. To this end, we propose a high-precision aircraft detection model specifically designed for such complex airport environments, combining gated convolutions with collaborative attention: SAGA-YOLO. Its network architecture is illustrated in Fig. 2.

3.1 Gated Backbone

3.1.1 Mathematical Model of SAR Speckle Noise

Unlike optical images, which are assumed to be subject to Gaussian noise, SAR generates coherent speckle noise due to coherent interference from radar echoes originating from multiple scatterers; this noise exhibits multiplicative non-Gaussian characteristics. This section introduces common mathematical models for SAR coherent speckle noise:

$$I(x,y) = S(x,y)\cdot u(x,y). \tag{1}$$

Here, (x,y) denotes spatial coordinates, $I(x,y)$ is the observed pixel intensity, $S(x,y)$ represents the true radar backscatter cross-section of the aircraft target or background clutter, and $u(x,y)$ is the coherent spot noise component. For an L-view SAR intensity image, the random noise term u follows a gamma distribution with a mean of 1, and its probability density function is expressed as:

$$p(u) = \frac{L^L u^{L-1}}{\Gamma(L)} \exp(-Lu), \quad u \geq 0. \tag{2}$$

Here, $\Gamma(\cdot)$ denotes the standard gamma function. In this case, the variance of the observed signal is:

$$\mathrm{Var}(I) = \frac{S^2(x,y)}{L}. \tag{3}$$

The formula shows that the noise power in SAR images is proportional to the signal intensity S^2; that is, the stronger the signal, the more intense the noise fluctuations, which exhibit severe non-Gaussian long-tail characteristics. When such multiplicative fluctuations pass through traditional static linear convolution layers, they are highly likely to amplify high-frequency noise signals and propagate them into deeper layers of the network, thereby obscuring weak-scattering structures of the aircraft, such as the wings or tail. This non-traditional assumption of fixed variance prevents the static-parameter convolution kernels in the original YOLO11 from adaptively handling this signal-related multiplicative interference.

3.1.2 The Principle of Noise Suppression

To address this challenge, inspired by the MambaOut architecture [23], this paper proposes a novel hierarchical gated convolutional backbone network to completely replace the feature extraction backbone of the original

YOLO11n. Its main component consists of four stacked Gated CNN Block modules. To maximize feature extraction and focus on high-level semantics, we employ an asymmetric depth stacking for the four-stage gated CNN modules, with the number of modules configured as [3, 3, 15, 3].

The structure of the Stem layer is shown in Fig. 3(a). As the feature input to the network, it performs initial downsampling of the input image using two layers of 3×3 convolutions with a stride of 2. This operation reduces the spatial resolution to one-fourth of the original, thereby reducing the computational load in subsequent stages, while mapping the input to a high-dimensional feature space, effectively filtering out a large amount of high-frequency background noise in SAR images early in the network. Subsequently, these feature maps sequentially pass through four network stages, with multiscale features further extracted between each stage via a downsample layer.

At each feature extraction stage, the core module consists of stacked gated convolutional neural network blocks, which are used for deep feature abstraction. The structure of the Gated CNN Block is shown in Fig. 3(c). It suppresses coherent speckle noise in SAR images through a dynamic channel-spatial gating mechanism; this suppression process is essentially a dynamic signal-to-noise ratio reweighting mechanism: Let the feature map entering this module be $X \in \mathbb{R}^{C\times H\times N}$. After initial linear channel expansion, the feature flow is split into a spatial convolution branch X_s and a gating branch g:

$$X_s = \Phi_{DW}\left(W_s X\right), \tag{4}$$

$$g = \mathrm{GELU}\left(W_g X\right). \tag{5}$$

Here, W_s and W_g represent the learnable linear projection weights for the two branches, respectively, while Φ_{DW} denotes deep separable convolution with large kernels, used to capture large-scale spatial context. GELU is a nonlinear activation function. Finally, dynamic modulation is achieved through element-wise multiplication, resulting in the purified feature Y:

$$Y = X_s \odot g. \tag{6}$$

For regions with high-signal-to-noise-ratio aircraft targets, aircraft—as strong metallic scatterers—exhibit deterministic, phase-correlated coherent reflections; their true scattering cross-sections $S(x, y)$ are large and highly correlated across channels. After undergoing a nonlinear mapping, the g branch assigns weights approaching 1 to these high-amplitude, structurally regular features. In this case, we have:

$$Y \approx X_s. \tag{7}$$

At this point, the gating is fully open, fully preserving the aircraft's geometric outline and key scattering point information.

For background clutter with a low signal-to-noise ratio and pure noise regions—such as airport runways and grassy areas—$S(x, y) \approx 0$, and the pixel values are almost entirely dominated by the randomly fluctuating Gamma noise term

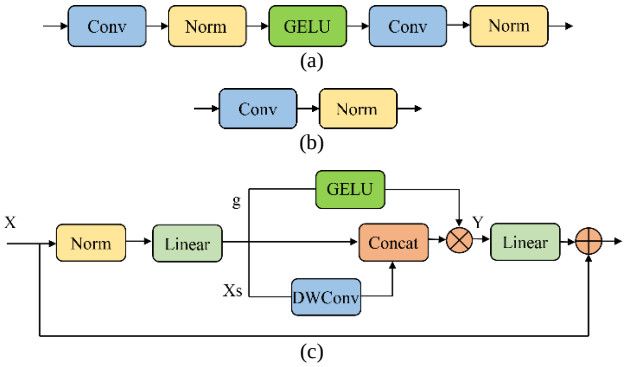


Fig. 3. Structures of the various modules in the improved backbone:(a) STEM Layer; (b) Downsample layer; (c) Gated CNN block.

$u(x, y)$, manifesting as isolated high-frequency spikes. Since these noise points lack spatial coherence and structural semantics, the gated branch outputs weights approaching 0 after deep spatial context capture and GELU negative clipping. At this point:

$$Y = X_s \odot g \approx 0. \tag{8}$$

Thus, the network implements a dynamic "information valve" at the physical level. It can adaptively modulate feature maps—strongly activating responses from high-scatter points representing the aircraft's core structure, while suppressing meaningless clutter from runways or buildings. Finally, through channel contraction and residual connections, the network outputs highly purified feature representations. Although stacking such Gated CNN blocks moderately increases the model's parameter count, the resulting "noise masking" and "feature purification" effects provide a decisive semantic foundation for improving the detection accuracy of subsequent models.

3.2 C2-CAS Module

To address the issue that the original C2PSA module is unable to effectively distinguish aircraft targets from background clutter when processing SAR images, this paper introduces the Convolutional Additive Self-attention (CAS) mechanism [24]. The core idea of the CAS module is to abandon the computationally intensive pairwise similarity calculations between tokens and instead construct global context through independent interactions across the spatial and channel dimensions, replacing 'multiplication' with 'addition'.

Given the input features $X \in \mathbb{R}^{C\times N}$, we first generate the query vector **Q**, the key vector **K** and the value vector **V** via independent linear transformations (Linear):

$$\mathbf{Q} = \mathbf{W}_q\mathbf{X}, \mathbf{K} = \mathbf{W}_k\mathbf{X}, \mathbf{V} = \mathbf{W}_v X. \tag{9}$$

Here, $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ are learnable weight matrices.

Subsequently, the CAS module does not directly compute the inner product of **Q** and **K**, but instead introduces a context mapping function $\Phi(\cdot)$. This function first passes through a spatial processing branch comprising depth-separable convolutions and a Sigmoid activation, and then enters

a channel processing branch comprising 1×1 convolutions and a Sigmoid activation, extracting local spatial information and channel semantic information from the features respectively. By passing **Q** and **K** through this mapping function and performing additive fusion, a novel similarity metric Sim(**Q**, **K**) is obtained:

$$\mathrm{Sim}(\mathbf{Q},\mathbf{K}) = \Phi(\mathbf{Q}) \oplus \Phi(\mathbf{K}). \tag{10}$$

Finally, the fused contextual information is subjected to a linear transformation $\Gamma(\cdot)$ and applied to the value vector **V** via element-wise multiplication ($\odot$), yielding the final attention output O:

$$O = \Gamma\left(\mathrm{Sim}(\mathbf{Q},\mathbf{K})\right) \odot \mathbf{V}. \tag{11}$$

To further capture the distinctive scattering characteristics of aircraft targets at the edge of the backbone network and suppress residual clutter, this paper combines the gradient advantages of the Cross Stage Partial (CSP) architecture with the focusing capabilities of CAS-ViT to construct the C2-CAS module, whose structure is shown in Fig. 4(a).

First, following the initial convolution, the input features are split into two parallel branches. One branch acts as an identity mapping, directly preserving the clean, high-level semantic information passed from the backbone network, thereby preventing the extracted deep features from degrading during repeated processing. Second, the other branch enters a stack of N CAS-PSA Blocks. As shown in the lower half of Fig. 4(a), within each block, the features first undergo spatial and channel-adaptive weighting via the CAS-ViT module. Unlike the complex matrix multiplication in traditional self-attention mechanisms, CAS-ViT extracts local contextual information simultaneously across both spatial and channel dimensions through a parallel dual-branch structure. Through this spatial-channel joint constraint and a smoothing additive attention mechanism, background noise that interferes with target recognition is gradually attenuated, while strong scattering features representing the aircraft's core components are repeatedly enhanced. Subsequently, the structural features of the aircraft target are further consolidated through residual connections and smoothing via two consecutive convolutional layers. Finally, the feature branch—refined multiple times by the CAS-PSA Block—is concatenated with the identity branch, which preserves the original semantic information, and feature dimension alignment is achieved through a terminal convolutional layer. Through this "two-pronged" mechanism, C2-CAS no longer focuses excessively on irrelevant noise at the edges of the aircraft but instead concentrates more attention on the aircraft's scattering structure. Its branch structure provides the model with raw, high-level semantic information; by fusing this information with the repeatedly refined feature information, the model significantly improves the accuracy of the final aircraft target localization and classification.

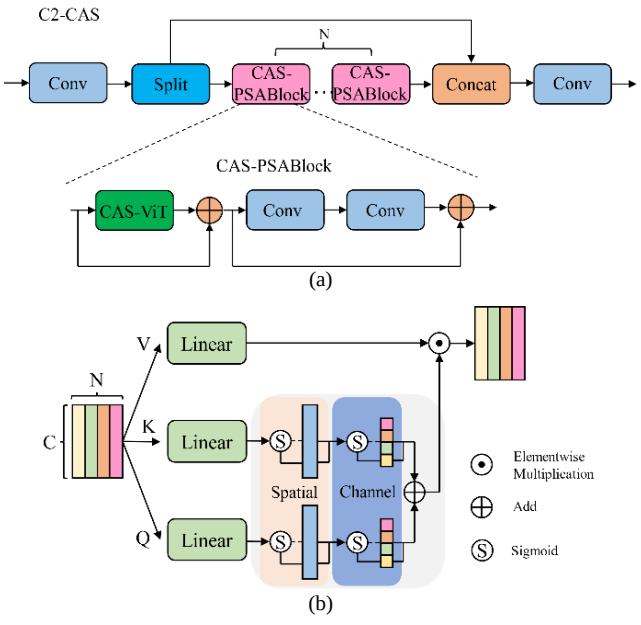


Fig. 4. Structure of the C2-CAS module.

3.3 Slim-neck Architecture

Data fusion is a key technique in machine learning for improving system accuracy and stability, and is typically categorized into early fusion and late fusion. Early fusion integrates multi-scale or multi-modal feature representations within a single network to enhance the model's representational capacity. Late fusion combines the outputs of multiple independent detectors or classifiers to produce a final consensus decision. Researchers often seek optimal fusion parameters under specific mathematical criteria, such as minimum mean squared error (LMSE) or minimum probability of error (MPE). In modern deep object detection networks, fusion typically occurs between multiscale feature maps within the network [25]. This paper introduces the Slim-neck architecture [26] to comprehensively reconstruct the neck network. This is a form of feature-level fusion performed between the backbone network and the detection head. In complex SAR scenarios, it effectively integrates shallow spatial details with deep semantic information, providing richer feature representations for subsequent target localization and recognition. Through the GSConv and VoV-GSCSP modules, this architecture significantly optimizes the feature extraction and fusion process. It moves away from the blind pursuit of absolute lightweight design and promotes cross-channel information flow between low-level and high-level features with exceptional efficiency.

The core of the Slim-neck architecture consists primarily of two fundamental components: the GSConv module and the VoV-GSCSP module.

Firstly, the GSConv module forms the basis for efficient feature transformation. Whilst traditional deep separable convolutions (DWConv) significantly reduce the number of parameters, they completely sever the inter-channel information connections within the input image during computation, resulting in a substantial decline in feature extraction capability. GSConv, however, combines standard convolution with DWConv. As shown in Fig. 5(a), the input features first undergo a standard convolution to halve the

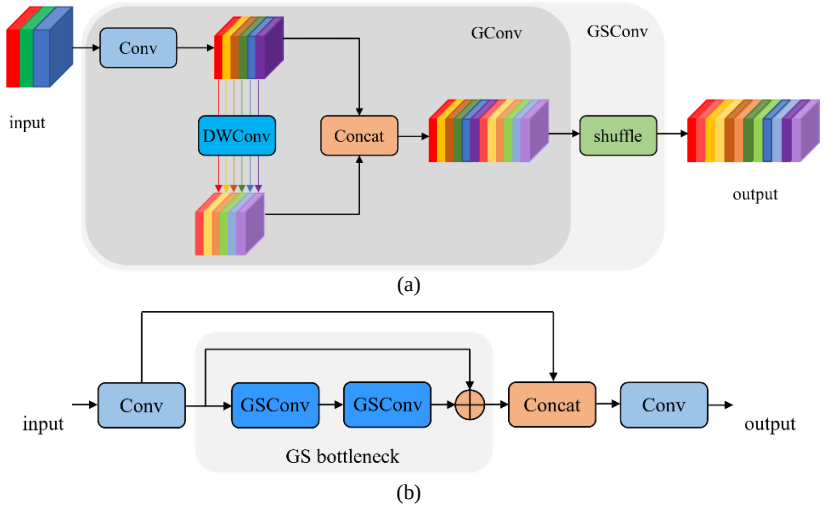


Fig. 5. Structure of Slim-neck: (a) GSConv; (b) VoV-GSCSP.

number of channels, after which these features are subjected to a DWConv operation. Finally, the two sets of features are concatenated, and a channel shuffle mechanism is introduced. The shuffle operation enables the dense information generated by the standard convolution to permeate uniformly into the sparse information generated by the DWConv, thereby restoring the hidden connections between channels.

As shown in Fig. 5(b), building upon the efficient feature transformation established by GSConv, Slim-neck utilizes the concept of One-Shot Aggregation (OSA) to design a cross-stage local network module, VoV-GSCSP, for performing deep feature fusion. In VoV-GSCSP, the input features are divided into two paths: the main path undergoes dimensionality reduction via convolution, followed by a GS bottleneck layer comprising multiple GSConv units in series, to perform deep-level non-linear feature extraction; the bypass branch acts as a skip connection, directly preserving the original, unprocessed high-resolution features. Finally, the features from these two paths are aggregated via a Concat operation and output through a standard terminal convolutional layer. This aggregation structure, based on the OSA concept, efficiently fuses features with different receptive fields and levels of abstraction in a 'one-shot' manner, significantly enhancing the network's ability to represent complex, multi-scale aircraft targets.

4. Experiments and Analysis

All experiments were conducted under identical conditions, using a 14th Gen Intel® Core™ i9-14900KF 2.4 GHz CPU and an NVIDIA GeForce RTX 4090D GPU with 24 GB of VRAM. The software environment comprised CUDA 12.2, Torch version 2.4.1, Python version 3.9.23, and the operating system was Ubuntu 24.04.1 LTS.

During the experiment, to ensure training stability and optimize GPU memory utilization, we set the batch size to 16. The input image size was set to 640 × 640, which represents a compromise given the 24 GB GPU memory limit when using a batch size of 16, while still preserving sufficient spatial detail for the multiscale aircraft targets in the dataset. The number of training epochs was set to 200 because, in our preliminary experiments, the validation loss and mAP metrics plateaued after approximately 150 epochs; the additional 50 epochs allowed the cosine annealing scheduler to reduce the learning rate to near zero, enabling fine-tuned convergence. The initial learning rate is set to 0.001, as we found this to be more stable than 0.01 and more efficient than 0.0001, based on our experimental results on the SAR validation set. In the data augmentation strategy, fliplr is set to 0.5 to horizontally flip training images with a 50% probability, thereby enriching the azimuth diversity of aircraft targets in SAR images and enhancing the model's generalization ability for targets facing different directions. Additionally, we set flipud to 0.2 to perform vertical flipping with a 20% probability, further expanding the distribution of training samples and enhancing the model's adaptability to variations in SAR image scattering characteristics. The parameters are shown in Tab. 1.

Parameter	Configuration
Image size	640 × 640
Batchsize	16
Epoch	200
Learning rate	0.001
Fliplr	0.5
Flipud	0.2

Tab. 1. Environmental parameter settings.

4.1 Introduction to the Dataset

The experiment utilized the publicly available dataset SAR-AIRcraft-1.0 [27] for model training and evaluation. This dataset is a specialized public dataset designed for aircraft target detection in SAR imagery. It was constructed to capture the scattering characteristics and distribution patterns of various aircraft types in airport scenarios, and is one of the standard datasets in the field of SAR remote sensing target detection. The dataset utilizes single-channel grey-

scale SAR images for annotation. Due to their metallic composition, aircraft targets exhibit strong radar scattering characteristics, appearing as bright regions in the images and forming a distinct contrast with airport runways and the background terrain. The images in this dataset are sourced from the Gaofen-3 satellite and comprise 4,368 images and 16,463 aircraft target instances. They cover seven categories: A220, A320/321, A330, ARJ21, Boeing 737, Boeing 787 and 'other', encompassing the differences in size, shape and scattering characteristics of mainstream civil aircraft. It is currently the high-resolution dataset with the widest range of aircraft types. The dataset images are available in resolutions of 800 × 800, 1000 × 1000, 1200 × 1200 and 1500 × 1500. The samples include aircraft targets at different azimuths, scales and densities, with scenes covering typical airport areas. This enables thorough validation of object detection models' capabilities regarding multi-pose, multi-class SAR aircraft targets, making it well-suited for relevant object detection tasks.

SADD (SAR Aircraft Detection Dataset) [28] is a dedicated SAR image dataset for aircraft target detection, released by the Remote Sensing Laboratory at Huazhong University of Science and Technology (hust-rslab). First made public in 2022, the data originates from Germany's TerraSAR-X satellite (X-band, HH polarization mode) and includes images at various resolutions—0.5 m, 1 m and 3 m—to simulate detection challenges under different imaging conditions. The dataset comprises a total of 2,966 images containing 7,835 aircraft targets, with each image measuring 224 × 224 pixels. The targets are situated in real-world scenarios such as airports and aprons, featuring various background noise elements including runways, buildings and vegetation. The dataset covers a range of aircraft parking angles and orientations, as well as partial occlusion scenarios, closely approximating real-world application scenarios. This dataset provides a standardized experimental benchmark for the field of aircraft detection in SAR imagery, and we utilize it for experiments assessing model generalization capabilities. The dataset is divided into training, test and validation sets in a 7 : 2 : 1 ratio. The distribution of aircraft bounding box sizes across the two datasets is shown in Fig. 6.

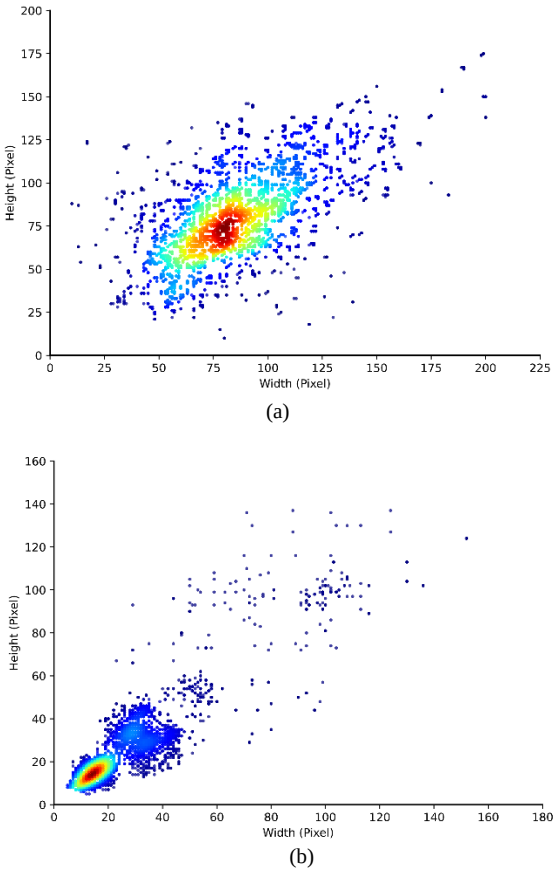


Fig. 6. Distribution of aircraft dimensions across two datasets: (a) SAR-AIRcraft-1.0; (b) SADD.

4.2 Evaluation Metrics

To assess the effectiveness of the model, we use precision (*P*), recall (*R*), average precision (*AP*), mean average precision (mAP), F1 score, frames per second (FPS) and the number of parameters as evaluation metrics. The formula for precision is given in (12), and the formula for recall is given in (13). In these equations, TP, FP, TN and FN represent, respectively, true positives (samples predicted as positive by the model), false positives (samples predicted as positive by the model that are actually negative), true negatives (samples predicted as negative by the model that are actually negative), and false negatives (samples predicted as negative by the model that are actually positive).

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{12}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{13}$$

The formulas for Average Precision (AP) and Mean Average Precision (mAP) are shown in (14) and (15), respectively. The value of AP is the area under the Precision-Recall curve, whilst mAP is the average of the AP values across all categories.

$$\mathrm{AP} = \int_0^1 P(R)\,\mathrm{d}R, \tag{14}$$

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i. \tag{15}$$

F1 score is a key metric used in object detection tasks to comprehensively evaluate a model's precision and recall; it ranges from 0 to 1, with higher values indicating superior overall performance in terms of both detection accuracy and completeness.

$$\mathrm{F1} = 2 \times \frac{P \times R}{P + R}. \tag{16}$$

FPS, or frames per second, is a key metric for measuring a model's inference speed and real-time performance; a higher value indicates that the model processes images

more quickly. In the formula, N represents the total number of image frames processed by the model, and T represents the total processing time (in seconds):

$$\mathrm{FPS} = \frac{N}{T}. \tag{17}$$

4.3 Comparative Experiments

To validate the superiority and effectiveness of the various improved modules proposed in this paper, we conducted comparative experiments between the C2-CAS module and other attention mechanism-based improvement methods, as well as between the Gated backbone and other mainstream visual backbone networks. All experiments were conducted on the SAR-AIRcraft-1.0 dataset.

Firstly, to validate the effectiveness of integrating the CAS attention mechanism into the C2PSA module, we conducted a comparison by applying current mainstream attention mechanisms to the C2PSA module. The control group included Efficient Multi-Scale Attention (EMA) [29], Discrete Cosine Transform Attention (DCT Attention) [30], Deformable Attention (DAttention) [31] and the Multi-scale Attention Block (MAB) [32]. The results are shown in Tab. 2.

Although C2-CAS has slightly more parameters than the other models under comparison, and its precision is only 0.1% lower than that of the MAB model, it demonstrates a clear lead in key evaluation metrics such as recall, mAP50 and mAP50-95, achieving 81.7%, 87.7% and 59.1% respectively. In particular, regarding recall, C2-CAS outperformed the high-performing EMA by 0.7%, effectively mitigating the issue of missed detections for targets with blurred, scattered structures or those in complex backgrounds. This demonstrates that a minor increase in the number of parameters has yielded a significant improvement in the model's overall detection performance, with C2-CAS achieving the optimal balance between high accuracy and high efficiency.

The feature extraction network of the original YOLO11 is unable to dynamically focus on key aircraft structural information when dealing with complex SAR airport environments; to address this, this paper constructs a Gated backbone based on a gating mechanism. To investigate the superiority of the Gated backbone in feature extraction, this comparative experiment is proposed. The backbone networks included in the comparison are: the PoolFormer [33], the Pyramid Vision Transformer v2 (Pvt_v2) [34], the large-kernel convolutional network UniRepLKNet [35], and the shift-window-based Swin Transformer [36]. The results are shown in Tab. 3.

Method	P/%	R/%	mAP50 /%	mAP50-95/%	Params/M
EMA	87.0	81.0	86.9	58.9	**2.53**
DCT	87.8	80.1	87.2	58.6	2.60
DAttention	86.9	79.1	86.0	58.2	2.61
MAB	**88.5**	79.4	86.9	58.5	2.59
CAS (Ours)	88.4	**81.7**	**87.7**	**59.1**	2.69

Tab. 2. Results of comparative experiments for the C2-CAS module.

Model	P/%	R/%	mAP50/ %	mAP50-95/%	Params/M
PoolFormer	83.6	79.2	85.4	57.9	13.29
Pvt_v2	85.9	83.1	87.4	60.7	15.39
UniRepLKNet	88.0	82.9	87.9	61.0	12.45
Swin Transformer	86.4	77.0	84.9	56.7	29.70
Gated backbone(Ours)	**89.5**	**83.9**	**88.3**	**63.7**	**10.80**

Tab. 3. Comparative experiments on improvements to the Gated Backbone architecture.

As can be seen from the experimental results, the Gated backbone achieved precision, recall, mAP50 and mAP50-95 of 89.5%, 83.9%, 88.3% and 63.7% respectively, outperforming all other backbone models. More notably, compared to other cumbersome backbone networks based on transformers or large-kernel convolutions, the Gated backbone has only 10.80 M parameters, which is significantly lower than other backbone networks. This is because the Gated backbone utilizes a gating mechanism to achieve efficient information flow and filtering, enabling it to extract richer and more robust global features without substantially increasing the number of parameters. The experimental results fully demonstrate that the Gated backbone, through the stacking of four Gated CNN Block modules, does not merely involve a simple stacking of parameters, but effectively and accurately extracts aircraft features tailored to the characteristics of SAR imagery.

4.4 Ablation Experiments

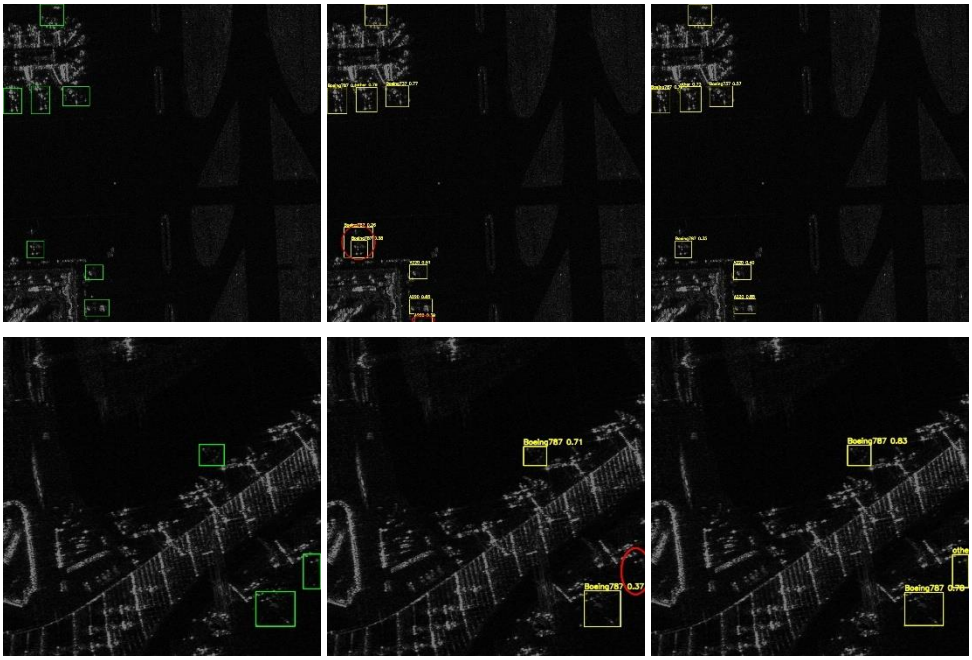
To further validate the effectiveness of each component within the improved network model, we designed the ablation experiments shown in Tab. 4 to analyze the practical performance of different improved modules on the SAR-AIRcraft-1.0 dataset. The ablation experiments employed the same evaluation metrics as the comparative tests, namely precision, recall, mean average precision (mAP50), mAP50-95, and the number of model parameters.

As shown in Tab. 4, the introduction of the Gated Backbone has effectively enhanced the model's detection capabilities, with mAP50 and the more stringent mAP50-95 increasing by 1.5% and 5.1%, respectively. Although the number of parameters has increased from 2.59 M to 10.8 M, this moderate and necessary investment in 'heavyweight' parameters is central to overcoming the SAR noise bottleneck and achieving robust semantic feature extraction. Relying solely on the simple convolutional stacking of the original model would make such high-intensity noise-resistant feature extraction fundamentally unachievable.

Building on this, we have replaced the C2PSA module in the original YOLO11 model with the C2-CAS module. Through the Channel Attention Scaling mechanism, the C2-CAS module can precisely capture the strong scattering centers characteristic of aircraft, such as wings, engines and tail surfaces, whilst filtering out clutter signals masquerading as targets. Following the addition of this module, the model's mAP50 improved to 88.6%, and mAP50-95 to 64.1%. The C2-CAS module adds only a negligible 0.1M parameters, yet plays a crucial role in 'feature refocusing'.

Gated backbone	C2-CAS	Slim-neck	P/%	R/%	mAP50/%	mAP50-95/%	Params/M
×	×	×	87.4	80.9	86.8	58.6	**2.59**
√	×	×	89.5	**83.9**	88.3	63.7	10.8
√	√	×	89.6	82.8	88.6	64.1	10.9
×	√	×	88.4	81.7	87.7	59.1	2.69
√	×	√	89.0	83.7	88.5	62.5	11.0
×	√	√	90.5	80.6	88.4	60.4	2.6
√	√	√	**91.0**	83.8	**89.8**	**64.4**	11.1

Tab. 4. Ablation experiment.



(a) (b) (c)

Fig. 7. Comparison of actual detection results: (a) Original image; (b) YOLOv11 detection results; (c) SAGA-YOLO detection results.

Finally, we upgraded the feature fusion architecture within the network to Slim-neck to address the issue of insufficient fusion between the abundant deep semantic information extracted from the backbone and shallow-level details. Experimental results show that the introduction of Slim-neck increased the model's precision from 89.6% to 91.0%, whilst achieving an mAP50 of 89.8%. It is worth noting that, as an efficient architecture, Slim-neck significantly enhances the model's performance whilst imposing virtually no additional computational burden.

We also compared detection performance in real-world scenarios; the experimental results are shown in Fig. 7. As can be seen, compared to the baseline model, SAGA-YOLO reduces false positives caused by background noise while correctly identifying even very small aircraft with relatively indistinct features. These improvements not only enhance the network's ability to detect multiscale objects but also effectively mitigate the negative impact of complex backgrounds on detection accuracy.

4.5 Comparison with Other Models

To comprehensively evaluate the detection performance of the SAGA-YOLO model under challenging SAR conditions, we conducted comparative experiments between this model and current mainstream classical object detection networks on the SAR-Aircraft-1.0 dataset, and validated the model's generalization capabilities on the SADD dataset. The experiments not only employed precision, recall, mean average precision (mAP50) and F1-score as performance evaluation metrics, but also introduced model parameters and frames per second (FPS) as key indicators reflecting the model's memory requirements and computational complexity. To analyze the variability in the experimental results, we conducted five Monte Carlo simulations for each comparison model. All metrics are based on the average of these five independent simulations, and the standard deviation (Std) of the key metric, mAP, is listed in the table.

4.5.1 SAR-AIRcraft-1.0 Dataset

For the purposes of analysis, the models used in this comparative experiment can be divided into three groups: the first group comprises the classic models SSD [15] and Faster R-CNN [12], along with the heavy-duty model RT-DETR [37]; the second group consists of the lightweight YOLO [14] model with a very small number of parameters; and the third group comprises YOLO models of medium parameter size. The results of the comparison are shown in Tab. 5.

As shown in the table, SAGA-YOLO outperforms other models on all three metrics—P, R, and mAP50. Compared to mid-sized models in the same series, the SAGA-YOLO model has only about half the number of parameters, yet it still maintains a slight lead in all detection metrics. Even the mid-sized YOLO11m from the same generation, with 20.05M parameters—far more than SAGA-YOLO—achieves an mAP50 of 87.7%, which is 2.1 percentage points lower. Furthermore, SAGA-YOLO achieves an inference speed of 232.56 FPS. Although this falls short of the latest YOLO11m and YOLO26m, its accuracy, recall, mAP50, and other metrics are superior to both of those versions, enabling efficient recognition while maintaining high accuracy.

To verify the statistical reliability of the proposed improvement, we performed a Welch's t-test on SAGA-YOLO and the baseline YOLO11n. On the SAR-AIRcraft-1.0 dataset, SAGA-YOLO achieved an average mAP50 of 89.8 ± 0.87%, while YOLO11n achieved 86.8 ± 0.87%, indicating a statistically significant improvement of 3.0%. Five independent experiments yielded $p = 0.0003$, Cohen's $d = 3.45$, and a 95% confidence interval of [1.73%, 4.27%]. The large effect size and strictly positive confidence interval confirm that this improvement is not only statistically significant but also meaningful in practical applications.

4.5.2 SADD Dataset

To further validate the generalization capability and robustness of the proposed model in the task of detecting aircraft targets using SAR, this paper conducted more extensive comparative experiments on the SADD dataset. In these generalization experiments, the classic two-stage detection model Faster R-CNN [12], the single-stage detection model SSD [15], and the currently mainstream YOLO [14] series of models were selected as benchmark models. The detailed results of the comparative experiments are shown in Tab. 6.

As can be clearly seen from the table, although traditional SSD and Faster R-CNN models achieved a certain level of detection accuracy on the SADD dataset, their recall metrics still lag significantly behind those of the YOLO series models. Compared to the advanced YOLO series algorithms, SAGA-YOLO also demonstrates outstanding overall performance. Its recall is only slightly lower than that of YOLOv8m, which has twice the number of parameters, while its accuracy and mAP50 outperform all other models. Furthermore, the model proposed in this paper achieves an inference speed of 357.14 FPS, which is significantly faster than heavyweight models such as YOLO11m, YOLOv8m, and YOLOv5m. Figure 8 shows a comparison of actual detection results on the SADD dataset. As shown in the figure, compared to the unmodified original YOLO11 model, SAGA-YOLO effectively reduces both false negatives and false positives.

On the SSDD dataset, although the baseline had already reached a high level, SAGA-YOLO still achieved a stable performance improvement. While the improvement exhibits a large effect size (Cohen's $d = 1.01$), indicating that our model performs better in practical applications, the p-value of 0.15 suggests that the improvement is limited due to the ceiling effect of the dataset.

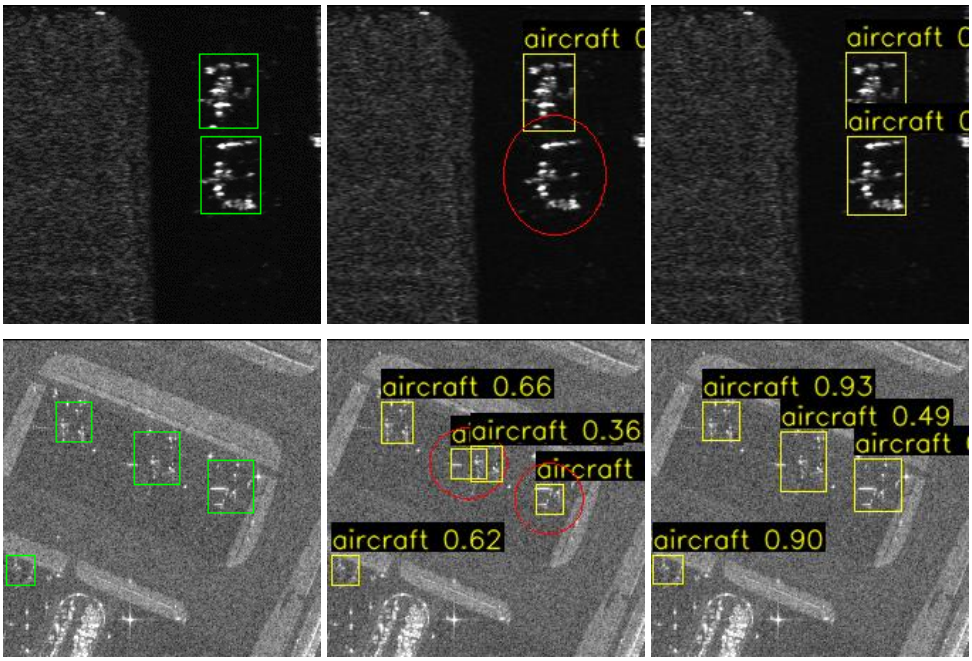
In summary, the model proposed in this paper not only performs excellently on the SAR-Aircraft-1.0 dataset but also demonstrates outstanding generalization performance on the SADD dataset, which features similar scenarios. It successfully achieves an optimal balance between detection accuracy, the number of parameters, and inference speed, maintaining stable, high-precision recognition of aircraft of various sizes in complex scenarios.

Models	P/%	R/%	mAP50/%±Std	Params/M	FPS	F1
SSD	51.2	40.4	51.2±0.89	26.3	36.77	0.452
Faster R-CNN	65.5	59.4	65.5±0.92	41.33	73.67	0.623
RT-DETR	88.6	82.2	85.3±0.56	31.9	135.14	0.853
YOLOv8n	87.9	80.5	88.6±0.77	3.1	**416.6**	0.840
YOLO11n	87.4	80.9	86.8±0.87	2.6	153.85	0.840
YOLO26n	80.2	66.2	74.1±0.75	**2.3**	158.73	0.725
YOLOv5m	90.2	81.9	89.2±0.90	25.1	91.74	0.858
YOLOv8m	90.3	82.7	88.9±0.60	25.86	212.77	0.863
YOLO11m	89.2	80.0	87.7±0.81	20.05	312.5	0.843
YOLO26m	83.2	76.4	82.5±0.93	20.35	370.37	0.796
Ours	**91.0**	**83.8**	**89.8**±0.87	11.1	232.56	**0.873**

Tab. 5. Comparison results for the SAR-AIRcraft-1.0 dataset.

Models	P/%	R/%	mAP50/%±Std	Params/M	FPS	F1
SSD	97.0	65.7	97.0±0.75	26.3	314.20	0.783
Faster R-CNN	97.9	84.6	97.9±0.66	41.3	89.21	0.908
YOLOv8n	96.8	97.3	98.9±0.10	3.1	625	0.970
YOLO11n	95.1	96.1	98.8±0.61	2.6	**909.09**	0.956
YOLO26n	91.5	93.1	97.1±0.98	**2.38**	625	0.923
YOLOv5m	96.9	97.7	98.9±0.42	25.1	277.7	0.973
YOLOv8m	96.4	**98.6**	98.8±0.30	25.86	250	0.975
YOLO11m	97.3	97.4	99.0±0.49	20.05	250	0.973
YOLO26m	93.9	96.1	98.2±0.76	20.35	384.62	0.950
Ours	**98.2**	98.3	**99.4**±0.58	11.1	357.14	**0.982**

Tab. 6. Comparison results for the SADD dataset.



(a) (b) (c)

Fig. 8. Comparison of actual detection results. (a) Original image; (b) YOLOv11 detection results; (c) SAGA-YOLO detection results.

5. Conclusions

This paper proposes SAGA-YOLO, a high-precision adaptive aircraft detection model specifically designed for complex SAR scenarios. Addressing the challenges inherent in SAR imagery—such as severe coherent speckle noise, strong ground structure clutter, and significant multi-scale variations in passenger aircraft targets—the model aims to achieve robust and high-precision target localization and recognition. To validate the model's effectiveness and generalization capability, we conducted comprehensive evaluations on the SAR-AIRcraft-1.0 and SADD large-scale high-resolution datasets, with test scenarios covering complex aprons, offshore environments, and a wide range of scales from small to wide-body aircraft.

Through a systematic analysis of the discrete scattering characteristics of aircraft and the distribution patterns of background noise in SAR images, this paper has undertaken a profound restructuring of the YOLO11 architecture. Specifically, we have constructed a hierarchical gated convolutional backbone network to dynamically focus on key scattering points and filter out low-level high-frequency noise; we innovatively designed a C2-CAS collaborative attention mechanism to precisely focus on and extract the core strong scattering features of the aircraft; simultaneously, we integrated an efficient multi-scale feature fusion architecture based on Slim-neck, which overcomes the interaction bottleneck between deep and shallow layers with an extremely high parameter-to-performance ratio.

However, whilst SAGA-YOLO achieves strong performance on existing SAR aircraft datasets, the gated backbone introduced to achieve optimal noise resistance and feature focusing results in a significant increase in the total number of parameters (11.1M) compared to the baseline model (2.6M). In terms of deployment, detection models of this order of magnitude can be further optimized using lightweight techniques such as structured pruning, knowledge distillation, and model quantization, and have already demonstrated viable deployment potential on mainstream mid-sized drone edge computing platforms [38]. Therefore, an important direction for future work is to optimize the model for practical deployment and inference on edge computing platforms while maintaining high model performance. Furthermore, due to the complexity of airport scenes and the presence of numerous interferences, the strong background scattering can affect aircraft recognition, and aircraft structures are relatively more complex; the differences in SAR images between aircraft of various types are negligible. Consequently, we plan to introduce a multimodal fusion network in our subsequent work, combining SAR images with optical remote sensing images of the same aircraft type. By utilizing the rich textures and geometric features of optical images to compensate for the lack of visual semantics in SAR images, we aim to further improve the accuracy of SAR aircraft target detection.

Acknowledgments

This work was supported by the Innovation and Research Institute of Hebei University of Technology in Shijiazhuang, Science and Technology Cooperation Special Project of Shijiazhuang (Grant No. SJZZXC23001).

References

[1] YAO, W. Q., CAO, L., TIAN, S., et al. Dynamic frequency-aware feature enhancement network for synthetic aperture radar target detection (in Chinese). *Laser and Optoelectronics Progress*, 2026, vol. 63, no. 6, p. 1–15. DOI: 10.3788/LOP252076

[2] WANG, C., DING, H., ZHAO, M. Aircraft target detection algorithm for SAR images based on improved YOLOv8. In *2024 30th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*. Leeds (UK), 2024, p. 1–6. DOI: 10.1109/M2VIP62491.2024.10746028

[3] KE, Q., WU, Y., ZHAO, W., et al. SFG-Net: A scattering feature guidance network for oriented aircraft detection in SAR images. *Remote Sensing*, 2025, vol. 17, no. 7, p. 1–25. DOI: 10.3390/rs17071193

[4] LIU, W., WANG, H., DUAN, J., et al. Complex-scene SAR aircraft recognition combining attention mechanism and inner convolution operator. *Sensors*, 2025, vol. 25, no. 15, p. 1–15. DOI: 10.3390/s25154749

[5] WANG, X., XU, W., HUANG, P., et al. MSADNet: Multistage SAR aircraft target detection network. *IEEE Geoscience and Remote Sensing Letters*, 2025, vol. 22, p. 1–5. DOI: 10.1109/LGRS.2024.3521650

[6] WANG, C. S., LI, Y. Z., SHANG, Y. Z., et al. An anchor-free method for aircraft detection in SAR images based on density map. In *2024 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*. Zhuhai (China), 2024, p. 1–6. DOI: 10.1109/ICSIDP62679.2024.10868269

[7] FINN, H. M., JOHNSON, R. S. Adaptive detection mode with threshold control as a function of spatially sampled clutter-level estimates. *RCA Review*, 1968, vol. 29, no. 3, p. 414–465. Available at: https://www.worldradiohistory.com/ARCHIVE-RCA/RCA-Review/RCA-Review-1968-09.pdf

[8] ROHLING, H. Radar CFAR thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, 1983, vol. AES-19, no. 4, p. 608–621. DOI: 10.1109/TAES.1983.309350

[9] NOVAK, L. M., OWIRKA, G. J., NETISHEN, C. M. Performance of a high-resolution polarimetric SAR automatic target recognition system. *The Lincoln Laboratory Journal*, 1993, vol. 6, no. 1, p. 11–24. Available at: https://archive.ll.mit.edu/publications/journal/pdf/vol06_no1/6.1.2.polarimetricsar.pdf

[10] APATEAN, A., ROGOZAN, A., BENSRHAIR, A. SVM-based obstacle classification in visible and infrared images. In *2009 17th European Signal Processing Conference*. Glasgow (Scotland), 2009, p. 293–297. Available at: https://www.eurasip.org/Proceedings/Eusipco/Eusipco2009/contents/papers/1569192405.pdf

[11] FREUND, Y., SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*. Berlin (Germany), 1995, p. 23 to 37. DOI: 10.1007/3-540-59119-2_166

[12] REN, S., HE, K., GIRSHICK, R., et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 6, p. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031

[13] CAI, Z., VASCONCELOS, N. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, vol. 43, no. 5, p. 1483 to 1498. DOI: 10.1109/TPAMI.2019.2956516

[14] REDMON, J., DIVVALA, S., GIRSHICK, R., et al. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas (USA), 2016, p. 779–788. DOI: 10.1109/CVPR.2016.91

[15] LIU, W., ANGUELOV, D., ERHAN, D., et al. SSD: Single shot multiBox detector. In *Computer Vision – ECCV 2016*. Amsterdam (Netherlands), 2016, p. 21–37. DOI: 10.1007/978-3-319-46448-0_2

[16] YANG, Y. J., SINGHA, S., MAYERLE, R. A deep learning based oil spill detector using Sentinel-1 SAR imagery. *International Journal of Remote Sensing*, 2022, vol. 43, no. 11, p. 4287–4314. DOI: 10.1080/01431161.2022.2109445

[17] SHEN, Y. F., GAO, Q. DS-YOLO: A SAR ship detection model for dense small targets. *Radioengineering*, 2025, vol. 34, no. 3, p. 407 to 421. DOI: 10.13164/re.2025.0407

[18] MO, H., WU, J., XIA, H., et al. A lightweight, efficient, adaptive design of YOLOv5 for enhanced SAR ship detection. *Remote Sensing Letters*, 2025, vol. 16, no. 5, p. 549–559. DOI: 10.1080/2150704X.2025.2480761

[19] ZHU, L. J., CHEN, J. L., CHEN, J. Y., et al. DGSP-YOLO: A novel high-precision synthetic aperture radar (SAR) ship detection model. *IEEE Access*, 2024, vol. 12, p. 167919–167933. DOI: 10.1109/ACCESS.2024.3497314

[20] ROCHA, R. D. L., FIGUEIREDO, F. A. P. D. Enhancing YOLO-based SAR ship detection with attention mechanisms. *Remote Sensing*, 2025, vol. 17, no. 18, p. 1–32. DOI: 10.3390/rs17183170

[21] WANG, L. G., NING, Y. X., WANG, G. Q., et al. AMDC-YOLO: An adaptive multi-dimensional dynamic convolution approach for aircraft target detection. In *2025 10th International Conference on Computer and Communication System (ICCCS)*. Chengdu (China), 2025, p. 278–283. DOI: 10.1109/ICCCS65393.2025.11069621

[22] BAYRAKTAR, I., BAKIRCI, M. Attention-augmented YOLO11 for high-precision aircraft detection in synthetic aperture radar imagery. In *2025 27th International Conference on Digital Signal Processing and Its Applications (DSPA)*. Moscow (Russia), 2025, p. 1–6. DOI: 10.1109/DSPA64310.2025.10977903

[23] YU, W. H., WANG, X. C. MambaOut: Do we really need mamba for vision? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville (USA), 2025, p. 4484–4496. DOI: 10.1109/CVPR52734.2025.00423

[24] ZHANG, T. F., LI, L., ZHOU, Y., et al. CAS-ViT: Convolutional additive self-attention vision transformers for efficient mobile applications. arXiv preprint, 2024. DOI: 10.48550/arXiv.2408.03703

[25] SALAZAR, A., SAFONT, G., VERGARA, L., et al. Graph regularization methods in soft detector fusion. *IEEE Access*, 2023, vol. 11, p. 144747–144759. DOI: 10.1109/ACCESS.2023.3344776

[26] LI, H., LI, J., WEI, H., et al. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. arXiv preprint, 2022. DOI: 10.48550/arXiv.2206.02424

[27] WANG, Z. R., KANG, Y. Z., ZENG, X., et al. SAR-AIRcraft-1.0: High-resolution SAR aircraft detection and recognition dataset. *Journal of Radars*, 2023, vol. 12, no. 4, p. 906–922. DOI: 10.12000/JR23043

[28] ZHANG, P., XU, H., TIAN, T., et al. SEFEPNet: Scale expansion and feature enhancement pyramid network for SAR aircraft detection with small sample dataset. *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, 2022, vol. 15, p. 3365–3375. DOI: 10.1109/JSTARS.2022.3169339

[29] OUYANG, D. L., HE, S., ZHANG, G. Z., et al. Efficient multi-scale attention module with cross-spatial learning. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island (Greece), 2023, p. 1–5. DOI: 10.1109/ICASSP49357.2023.10096516

[30] JIANG, M., ZENG, P., WANG, K., et al. FECAM: Frequency enhanced channel attention mechanism for time series forecasting. 2022, *arXiv*: arXiv:2212.01209, p. 1–11. DOI: 10.48550/arXiv.2212.01209

[31] XIA, Z., PAN, X., SONG, S., et al. Vision transformer with deformable attention. *arXiv*, 2022, arXiv:2201.00520, p. 1–12. DOI: 10.48550/arXiv.2201.00520

[32] WANG, Y., LI, Y., WANG, G., et al. Multi-scale attention network for single image super-resolution. *arXiv*, 2024, arXiv:2209.14145, p. 1–11. DOI: 10.48550/arXiv.2209.14145

[33] YU, W. H., LUO, M., ZHOU, P., et al. MetaFormer is actually what you need for vision. *arXiv*: 2022, arXiv:2111.11418, p. 1–17. DOI: 10.48550/arXiv.2111.11418

[34] PENG, Y. P., CHEN, D. Z., SONKA, M. U-Net V2: Rethinking the skip connections of U-Net for medical image segmentation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. Houston (USA), 2025, p. 1–5. DOI: 10.1109/ISBI60581.2025.10980742

[35] DING, X. H., ZHANG, Y. Y., GE, Y. X., et al. UniRepLKNet: A universal perception large-kernel ConvNet for audio, video, point cloud, time-series and image recognition. *arXiv*: 2024, arXiv:2311.15599, p. 1–16. DOI: 10.48550/arXiv.2311.15599

[36] LIU, Z., LIN, Y. T., CAO, Y. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal (Canada), 2021, p. 10012–10022. DOI: 10.1109/ICCV48922.2021.00986

[37] ZHAO, Y., LV, W. Y, XU, S. L., et al. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle (USA), 2024, p. 16965–16974. DOI: 10.1109/CVPR52733.2024.01605

[38] SABAGHIAN, M., KEYVANRAD, M. A., MOGHADAMI, S. M. A novel compression framework for YOLOv8: Achieving real-time aerial object detection on edge devices via structured pruning and channel-wise distillation. *arXiv*: 2025, arXiv:2509.12918, p. 1–28. DOI: 10.48550/arXiv.2509.12918

About the Authors ...

Qing GUO was born in 1999. He received the B.S. degree in Electronic Information Engineering from Anhui Polytechnic University, Wuhu, China, in 2021. He is currently pursuing the M.S. degree in Electronic Information with Hebei University of Technology, Tianjin, China. His research interests include SAR image target recognition and machine learning.

Hongdong ZHAO (corresponding author) was born in 1968. He received the Ph.D. degree from Beijing University of Technology, Beijing, China, in 1998. He is currently a Full Professor with Hebei University of Technology, Tianjin, China. His research interests include semiconductor laser, computer vision, and machine learning. Dr. Zhao is a Senior Member of Chinese Optical Society and an Editorial Board Member of Laser Optoelectronics.

Wenjing WANG was born in 2002. She received the B.S. degree in Electronic Science and Technology from Hebei University, Baoding, China, in 2024. She is currently pursuing the M.S. degree in Electronic Information with Hebei University of Technology, Tianjin, China. Her research interests include deep learning and point clouds.

Wenkai ZHANG was born in 2002. He received the B.S. degree in Electronic Science and Technology from Hebei University, Baoding, China, in 2024. He is currently pursuing the M.S. degree in Electronic Information with Hebei University of Technology, Tianjin, China. His research interests include deep learning and computer vision.