CGAAN: CFAR-Guided Architecture-Adaptive Network for SAR Target Detection

Lingjuan Yu , Xinya Xiong , Xiaochun Xie , Miaomiao Liang * , Xiangchun Yu , Xuan Jiao , Wen Hong

Posted Date: 7 May 2026

doi: 10.20944/preprints202605.0337.v1

Keywords: constant false alarm rate (CFAR); deep learning; synthetic aperture radar (SAR); target detection; YOLOv8

 doi:10.20944/preprints202605.0337.v1

Article

CGAAN: CFAR-Guided Architecture-Adaptive Network for SAR Target Detection

Lingjuan Yu [1], Xinya Xiong [1], Xiaochun Xie [2], Miaomiao Liang [1,*], Xiangchun Yu [1], Xuan Jiao [1] and Wen Hong [3]

[1] Jiangxi Province Key Laboratory of Multidimensional Intelligent Perception and Control, School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China
[2] School of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000, China
[3] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100194, China
* Correspondence: liangmiaom@jxust.edu.cn

Highlights

What are the main findings?

- Dataset complexity is quantified under the CFAR principle by measuring the proportion of pixels whose locally adaptive thresholds exceed a predefined global reference.
- Based on the quantified dataset complexity, a structure-adaptive SAR target detection network is designed to flexibly adjust network architecture for diverse SAR scenarios.

What are the implications of the main findings?

- Dataset complexity is closely related to the required depth of multi-level feature fusion in SAR target detection. Deeper feature fusion is more beneficial for high-complexity datasets, while shallower fusion is sufficient for relatively low-complexity scenarios.
- The proposed CFAR-guided architecture-adaptive network (CGAAN) consistently outperforms representative detectors on SAR-Aircraft-1.0 and HRSID, demonstrating its effectiveness and stability across diverse SAR scenarios.

Abstract

Improving robustness across diverse SAR scenes remains a key challenge in deep learning-based SAR target detection. To address this issue, we propose a CFAR-guided architecture-adaptive network (CGAAN), which adjusts its network structure according to dataset complexity. Specifically, dataset complexity is quantified under the CFAR principle by computing the proportion of pixels whose locally adaptive thresholds exceed a predefined global reference, thereby reflecting background clutter and detection difficulty. Based on this indicator, an architecture-adaptive YOLOv8 is constructed with three key components. First, a lightweight representation-enhanced backbone integrating ResNet18 and a dilated convolutional spatial pyramid (DCSP) module is adopted to improve contextual representation while maintaining low model complexity. Second, a structure-adaptive neck (SAN) is further developed to regulate multi-level feature fusion according to dataset complexity. Third, a Complete Intersection over Union (CIoU)-modulated head (CMH) is developed to enhance classification-regression alignment and suppress clutter-induced responses. Experiments on SAR-Aircraft-1.0 and HRSID datasets indicate that deeper feature fusion benefits high-complexity datasets, whereas shallower fusion is sufficient for low-complexity scenarios. Moreover, the proposed CGAAN achieves superior performance over representative detectors, demonstrating its effectiveness and stability on SAR datasets with different scene characteristics.

Keywords: constant false alarm rate (CFAR); deep learning; synthetic aperture radar (SAR); target detection; YOLOv8

 doi:10.20944/preprints202605.0337.v1

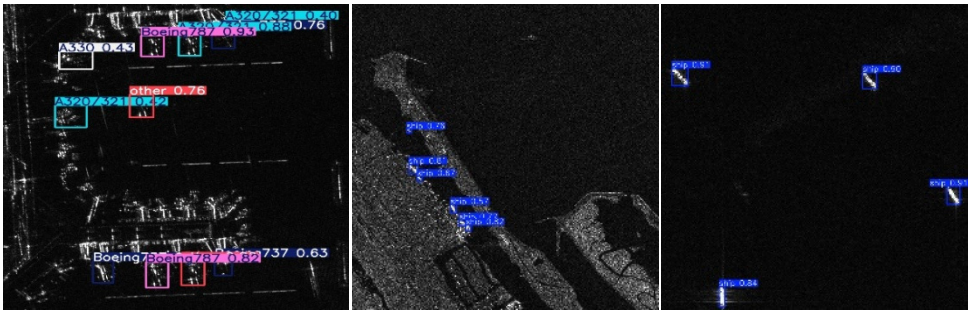
1. Introduction

Synthetic aperture radar (SAR) is an active microwave sensor that provides all-day and all-weather imaging capability as well as strong penetration ability. These advantages have enabled its widespread applications in disaster assessment, land-use monitoring, maritime surveillance, and military reconnaissance. Among various SAR image interpretation tasks, target detection plays a fundamental role by automatically localizing and identifying objects of interest, such as aircraft, ships, and vehicles, in SAR imagery [1]. However, due to the coherent imaging mechanism of SAR, the acquired images are inherently affected by multiplicative speckle noise, complex scattering behavior, and heterogeneous background clutter, which significantly reduce the separability between targets and background clutter. These characteristics make SAR target detection still challenging.

Early SAR target-detection methods mainly relied on statistical modeling to characterize the differences between targets and backgrounds. Constant false alarm rate (CFAR) detectors were among the most widely used approaches [2]. CFAR estimated local background statistics and adaptively determined detection thresholds to maintain a constant false alarm probability. Owing to their low computational complexity and clear physical interpretability, CFAR-based methods have been extensively deployed in practical systems. Nevertheless, their performance was fundamentally constrained by assumptions on local clutter distribution. In complex environments such as near-shore, port, or urban areas, clutter heterogeneity and model mismatch degraded detection reliability, limiting the generalization capability of CFAR-based detectors.

In recent years, deep learning-based SAR target detection has gradually become the dominant paradigm. By learning hierarchical nonlinear feature representations end-to-end, deep SAR detectors significantly outperform traditional approaches based on handcrafted features in both representation capability and detection accuracy. From the perspective of detection pipelines, existing methods can be broadly categorized into two-stage detectors [3–5] and one-stage detectors [6–11]. From an architectural perspective, most high-performance SAR target detection methods are still constructed on convolutional neural networks (CNNs) [12–17], benefiting from their strong inductive biases and efficient local feature modeling. More recently, CNN-Transformer hybrid architectures are more commonly adopted, in which CNNs extract multi-scale local features, and Transformers capture long-range dependencies, thereby improving robustness in complex SAR scenes while maintaining computational efficiency [18–20].

Despite these advances, the performance of deep SAR detectors remains highly dependent on the compatibility between network architecture and dataset characteristics. In real cases, SAR datasets often exhibit substantial variations in scene complexity, target density, and clutter distribution. As illustrated in Figure 1(a), SAR-Aircraft-1.0 [21] is a fine-grained aircraft detection dataset characterized by complex backgrounds. In contrast, the HRSID dataset [22], shown in Figure 1(b), is dedicated to ship detection without fine-grained classification, and most of the images correspond to offshore scenes with relatively simple backgrounds. Such differences in target granularity and clutter intensity imply that a unified architecture is hard to achieve optimal performance across different datasets. The optimal feature extraction depth and fusion strategy should be varied across datasets with distinct levels of complexity.



(a) (b) (c)

 doi:10.20944/preprints202605.0337.v1

Figure 1. Two datasets with different complexities. (**a**) SAR-AIRcraft-1.0 dataset. (**b**) HRSID dataset (onshore). (**c**) HRSID dataset (offshore).

To address model adaptability, existing approaches mainly follow two directions. The first seeks to improve cross-domain generalization through domain adaptation or domain generalization techniques [23–26], which attempt to align feature distributions across different datasets. The second relies on neural architecture search (NAS) to automatically explore optimal structures within predefined search spaces [27]. The first one aims to develop a general-purpose model, whereas the second one focuses on a specialized model. However, these methods exhibit some shortcomings. Domain adaptation typically requires target-domain data, domain generalization requires strict assumptions about category-space alignment, and NAS entails substantial computational overhead and lacks interpretability.

In this work, a data-driven compromise model between general and specialized models is proposed. Inspired by the cell-averaging CFAR (CA-CFAR) principle [28], we first evaluate dataset complexity by computing the proportion of pixels whose local adaptive thresholds exceed a predefined global reference threshold. Then, an architecture-adaptive detection network is constructed based on YOLOv8, which can be adjusted to align model capacity with the characteristics and complexity levels of SAR data. The proposed network keeps the overall framework unchanged across datasets, while the local structures for feature fusion are adaptively determined based on dataset complexity. The primary contributions are summarized as follows:

(1) A CA-CFAR-based complexity metric is proposed to characterize dataset-level detection difficulty. For each dataset, local adaptive thresholds are estimated using the CA-CFAR detector, and the proportion of thresholds exceeding a predefined global reference threshold is used as a quantitative indicator of background clutter intensity. Based on this indicator, the complexity of different SAR datasets can be objectively compared.

(2) A lightweight representation-enhanced backbone is designed by integrating ResNet18 and a dilated convolution spatial pyramid (DCSP) module. ResNet18 is adopted for its suitability to SAR datasets with limited training samples. By employing cascaded dilated convolutions with shared parameters, DCSP enlarges the receptive field and strengthens contextual feature modeling while keeping the parameter overhead low.

(3) A structure-adaptive neck (SAN) is proposed to tailor multi-level feature fusion to dataset complexity. By adaptively adjusting the aggregation depth and cross-scale interactions, SAN facilitates more effective feature fusion across datasets with diverse background-clutter levels.

(4) A CIoU-modulated head (CMH) is designed to improve classification-regression alignment by reweighting predictions according to localization reliability, thereby emphasizing reliable samples, suppressing clutter-affected unreliable ones, and enhancing prediction consistency.

Experiments on SAR-Aircraft-1.0 and HRSID show that adapting feature-fusion depth to dataset complexity enables CGAAN to achieve superior and stable detection performance under different SAR scene characteristics.

2. Related Work

2.1. CFAR-Based SAR Target Detection

The CFAR represents a classical adaptive detection framework based on local background statistical modeling and threshold-based decision-making. By estimating clutter statistics within sliding reference windows and determining the detection threshold based on a predefined false-alarm rate, CFAR maintains stable false-alarm control under varying clutter power. However, its performance inherently relies on two critical assumptions: local background homogeneity and the absence of contamination from targets or outliers [29]. When these assumptions are violated in heterogeneous scenes, both false alarms and missed detections increase significantly.

 doi:10.20944/preprints202605.0337.v1

To address heterogeneous backgrounds, prior studies have primarily focused on improving the robustness of background estimation. Robust statistical CFAR variants employed truncated or censored samples and redesigned parameter estimation strategies to mitigate interference [2]. Region- or superpixel-based CFAR schemes replaced pixel-wise sliding windows with homogeneous spatial units obtained through segmentation, thereby enhancing background purity and reducing estimation variance [30]. These improvements suggested that CFAR performance enhancement was primarily achieved through refined clutter modeling rather than fundamentally new detection mechanisms.

Recently, CFAR has been extended toward system-level implementation and integration with learning-based frameworks. GPU-oriented implementations reformulated classical CFAR operations into parallel tensor-based computations to support large-scale and real-time SAR processing [31]. In addition, CFAR-guided deep learning approaches exploited CFAR outputs as interpretable priors or proposal generators in complex scenarios [32]. Unlike these approaches, this work employs a simple CA-CFAR scheme not as a detector but as a statistical tool to quantify dataset-level complexity, thereby providing an explicit basis for complexity-aware architecture adaptation.

2.2. Objects in SAR Target Detection

Existing SAR target detection studies have mainly focused on aircraft and ships, whereas other target categories have received relatively limited attention. Most deep learning-based methods have been developed for specific target categories. Accordingly, studies on aircraft and ship detection have primarily addressed the challenges encountered in their respective detection scenarios.

For aircraft detection, the main challenges include discrete strong scattering, structural discontinuities, incomplete contours, scale variations, complex background clutter, airport-dependent spatial distributions, and fine-grained category ambiguity. To address structural fragmentation, scattering-point- and scattering-region-based models have been developed to restore structural integrity [33–36]. To improve scale robustness, multi-scale feature extraction and fusion, together with reformulation of the detection framework, have been adopted [37,38]. To suppress complex background clutter and speckle interference in airport scenes, saliency cues and contextual information have been incorporated to enhance target-background discrimination [34,36,38]. In addition, geospatial priors and probabilistic constraints have been utilized to characterize the aircraft-airport relationship for airport-aware detection [39]. For fine-grained aircraft detection, global structural cues and category relationship constraints have been introduced to improve discrimination performance [39].

For ship detection, the main challenges include small targets and weak scattering, scale variation, dense adjacency, arbitrary orientations, complex inshore clutter, speckle noise, and cross-domain shift. To address these issues, multi-scale feature extraction and fusion, shallow-detail enhancement, and detection head optimization have been introduced to improve small-target sensitivity and scale robustness [13]; anchor-free, key-point-based, and query-based detection methods have been adopted to alleviate target confusion in densely distributed scenes [10,13]; oriented bounding box representation and geometry-aware optimization have been developed to better characterize arbitrary orientations and elongated structures [15,16,20]; contextual information and attention mechanisms have been incorporated to suppress background clutters in complex inshore environments [15,17,19]; and feature alignment and domain adaptation have been explored to improve generalization across SAR datasets [23–25].

In fact, the above challenges in aircraft and ship detection exhibit both common and category-specific characteristics. Shared challenges include multi-scale variation, as well as speckle noise, background clutter, and target–background confusion. Meanwhile, category-specific challenges mainly stem from differences in scattering patterns, structural characteristics, and scene priors. Therefore, an effective detection framework should not rely on a fixed architecture, but instead balance shared representation capability with dataset-specific adaptability. That is, the network should maintain sufficient common modeling capacity for generic SAR detection while being able to

 doi:10.20944/preprints202605.0337.v1

adjust its structure to the distinct characteristics of different target categories and scenes. To this end, an architecture-adaptive network for SAR target detection is proposed.

3. Method

3.1. Overall Framework

The overall framework of CGAAN is illustrated in Figure 2. It consists of two sequential stages: dataset complexity assessment and architecture-adaptive detection. In the first stage, a CA-CFAR-based statistical analysis is performed on the training dataset to quantify clutter intensity and detection difficulty. The resulting complexity indicator provides a training-free and interpretable measure of dataset-level background heterogeneity. In the second stage, an architecture-adaptive detection model is constructed upon YOLOv8. According to the estimated dataset complexity, a predefined structural configuration is activated to align model capacity with scene characteristics. The selected architecture comprises three major components: a lightweight representation-enhanced backbone for feature extraction, a SAN for complexity-aware multi-scale feature fusion, and a CMH for reliability-aware classification-regression alignment. The detailed descriptions are presented in the following subsections.

3.2. CA-CFAR-Based Dataset Complexity Assessment

To quantitatively characterize dataset-level detection difficulty, a CA-CFAR-based statistical framework is adopted. Following the classical CA-CFAR detection principle, local background statistics are estimated from reference cells surrounding the cell under test (CUT), and an adaptive threshold is computed to meet a predefined false-alarm probability. Unlike conventional CFAR detection, which performs a binary decision for target presence, the proposed framework further exploits the spatial distribution of adaptive thresholds to assess dataset complexity. By introducing a common global intensity benchmark, the proportion of pixels whose local adaptive thresholds exceed this benchmark is calculated as a dataset complexity indicator. A higher proportion implies stronger background heterogeneity and greater detection difficulty.

Suppose a SAR image has size $W \times H$, where W and H denote the width and height, respectively. For a CUT located at position (w, h) ($1 \le w \le W$, $1 \le h \le H$), a two-dimensional sliding window is defined and divided into three regions: the CUT, guard cells, and reference cells. The CUT lies at the center of the window. Guard cells form a surrounding region adjacent to the CUT to prevent target leakage into background estimation. Reference cells are located outside the guard region and are assumed to contain pure background clutter.

If the total number of reference pixels is M, the local average clutter power Z is computed as,

$$Z(w,h) = \frac{1}{M}\sum_{m=1}^{M} X_m \tag{1}$$

where X_m denotes the amplitude value of the m-th (m=1, 2, …, M) reference cell.

Under a predefined false alarm probability P_{fa}, the adaptive threshold for the CUT is calculated as,

$$T(w,h) = Z(w,h)\cdot\gamma \tag{2}$$

where the scaling factor γ is determined by,

$$\gamma = M\left(P_{fa}^{-1/M} - 1\right). \tag{3}$$

From (1)-(3), the adaptive threshold is statistically consistent with the assumed clutter distribution under the constant false alarm rate constraint.

To evaluate dataset complexity, a global reference threshold T_{global} is introduced as a unified intensity benchmark. The complexity indicator is defined as the ratio of the number of pixels with $T \ge T_{global}$ to the total number of pixels in the dataset. It can be written by,

 doi:10.20944/preprints202605.0337.v1

$$Ratio = \frac{\sum_{i=1}^{N_D}\sum_{w=1}^{W}\sum_{h=1}^{H}\mathbb{I}\left(T(w,h) \geq T_{global}\right)}{W \times H \times N_D} \tag{4}$$

where N_D represents the number of images in the dataset, and $\mathbb{I}(\cdot)$ is the indicator function.

According to (4), a larger ratio indicates greater clutter intensity and higher background heterogeneity. Based on the estimated complexity indicator, two structural modes are predefined: Mode S (shallow fusion) for low-complexity datasets and Mode D (deep fusion) for high-complexity datasets. This method enables objective and training-free complexity estimation and provides a principled basis for subsequent architecture adaptation.

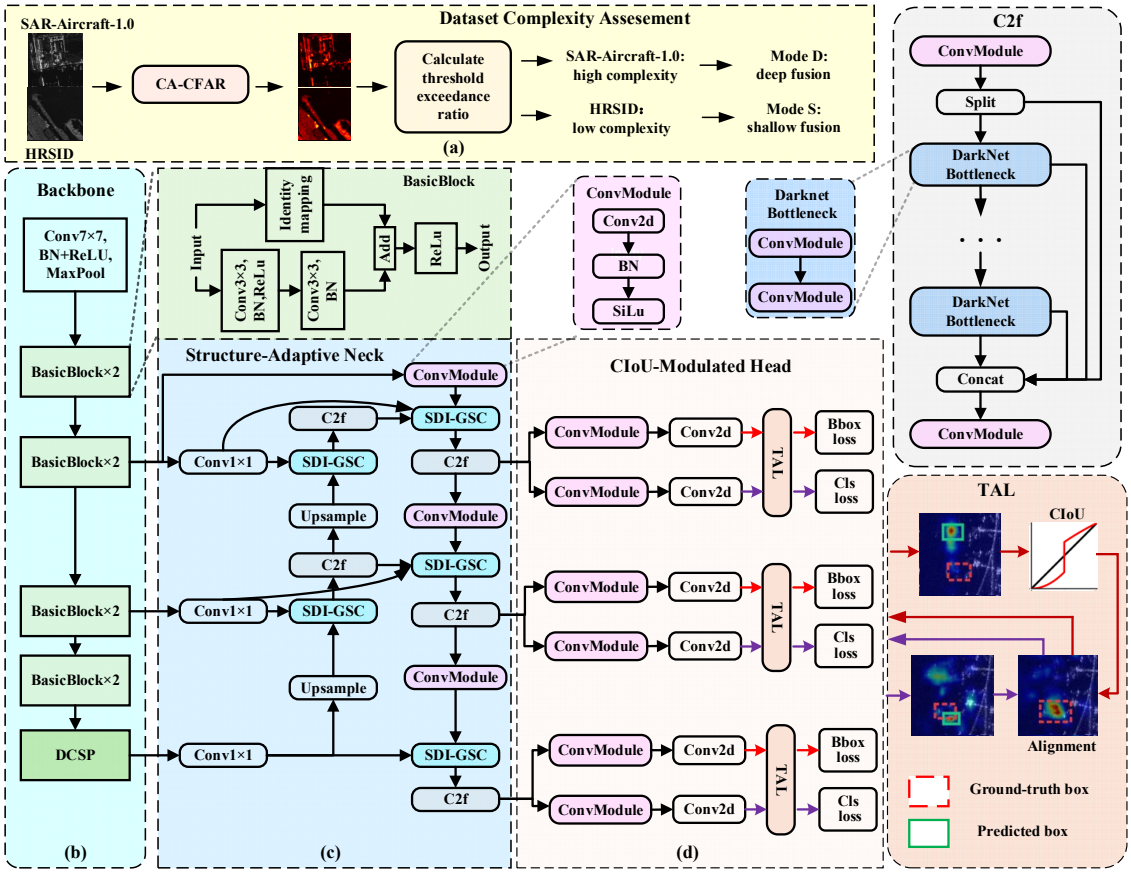


Figure 2. The overall framework of CGAAN. (**a**) CA-CFAR-based complexity assessment. (**b**) Backbone. (**c**) SAN. (**d**) CMH.

3.3. Lightweight Representation-Enhanced Backbone

A lightweight backbone integrating ResNet18 with a DCSP module is adopted for feature extraction. ResNet18 is selected for its moderate depth and stable optimization behavior, which are well-suited to SAR datasets with limited training samples. Compared with deeper architectures, it alleviates the risk of overfitting while preserving representation capacity. Besides, the DCSP module, derived from the original spatial pyramid pooling-fast (SPPF) module in YOLOv8, is integrated into the backbone to enhance contextual perception without introducing excessive parameters. Unlike SPPF, which enlarges the receptive field through fixed pooling operations, DCSP employs cascaded dilated convolutions with shared weights to achieve learnable receptive-field expansion.

The detailed structures of DCSP and SPPF are shown in Figure 3(a) and (b), respectively. The DCSP module applies a standard 1×1 convolution to reorganize channel information, followed by three cascaded 3×3 dilated convolution operations with different dilation rates. The intermediate feature maps can be expressed as,

 doi:10.20944/preprints202605.0337.v1

$$\begin{cases} Y_0 = Conv_{1\times1}(X) \\ Y_1 = Conv_{3\times3}^{d=1}(Y_0, W_{share}) \\ Y_2 = Conv_{3\times3}^{d=3}(Y_1, W_{share}) \\ Y_3 = Conv_{3\times3}^{d=5}(Y_2, W_{share}) \end{cases} \tag{5}$$

where X denotes the input feature maps, $Conv(\cdot)$ represents a convolution operation, d is the dilation rate, and W_{share} denotes the shared convolution parameters across the cascaded layers.

The final output of DCSP can be expressed by,

$$Y_{final} = Conv_{1\times1}(Concat(Y_0, Y_1, Y_2, Y_3)) \tag{6}$$

where $Concat(\cdot)$ denotes channel-wise concatenation.

The DCSP design provides two primary advantages. First, progressive dilation enlarges the effective receptive field in a learnable, hierarchical manner, which is beneficial for modeling anisotropic scattering patterns and contextual dependencies in SAR imagery. Second, weight sharing across the dilated convolutions reduces parameter redundancy and introduces implicit regularization, encouraging scale-consistent feature learning while maintaining computational efficiency.

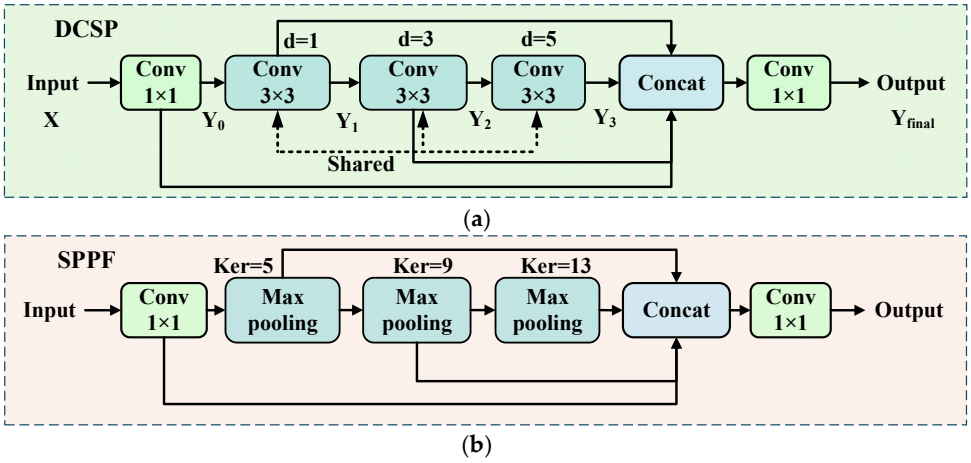


Figure 3. The structures of DCSP and SPPF. (**a**) DCSP. (**b**) SPPF.

3.4. Structure-Adaptive Neck

To align feature aggregation depth with dataset complexity, SAN is designed for feature fusion. In the original YOLOv8, the feature pyramid network and path aggregation network (FPN-PAN) use fixed multi-scale fusion pathways, with feature interaction patterns remaining unchanged regardless of scene characteristics. However, in SAR imagery, the appropriate fusion depth is influenced by factors such as clutter intensity and target scale distribution. Datasets with strong clutter tend to benefit from deeper cross-scale contextual modeling, whereas simpler scenes may favor shallower feature fusion to preserve local discriminative details. This adaptive design helps mitigate over-smoothing.

Inspired by the information flow of bidirectional FPN (BiFPN), SAN introduces a complexity-aware cross-layer fusion mechanism. Instead of using linear weighted summation as in BiFPN, SAN integrates semantic and detail infusion (SDI) [40] and grouped shuffle convolution (GSC) [41] to enhance both feature expression and inter-channel information exchange. The structure of SDI-GSC is shown in Figure 4(a), and the simplified version is shown in Figure 4(b). Before fusion, feature maps from different scales are first transformed into a unified spatial resolution through up-sampling, identity mapping, or down-sampling operations. Suppose that the features to be fused originate from K branches, and the feature map of the k-th (k=1, 2, …, K) branch is denoted as X_k. To unify all features to the spatial resolution of the r-th ($1 \le r \le K$) branch, the transformed feature map can be expressed by,

 doi:10.20944/preprints202605.0337.v1

$$Y_{kr} = \begin{cases} D(X_k) & k < r \\ I(X_k) & k = r \\ U(X_k) & k > r \end{cases} \tag{7}$$

where $D(\cdot)$ represents a down-sampling operation; $I(\cdot)$ represents an identity mapping operation; $U(\cdot)$ represents an up-sampling operation.

After scale unification, GSC is employed to enhance inter-channel interaction, and the detailed structure is shown in Figure 4(c). It combines standard convolution and depth-wise convolution to generate both intrinsic and ghost features at low computational cost, followed by a channel shuffle operation to further enhance cross-channel information exchange. This design improves parameter efficiency while preserving strong representation capability.

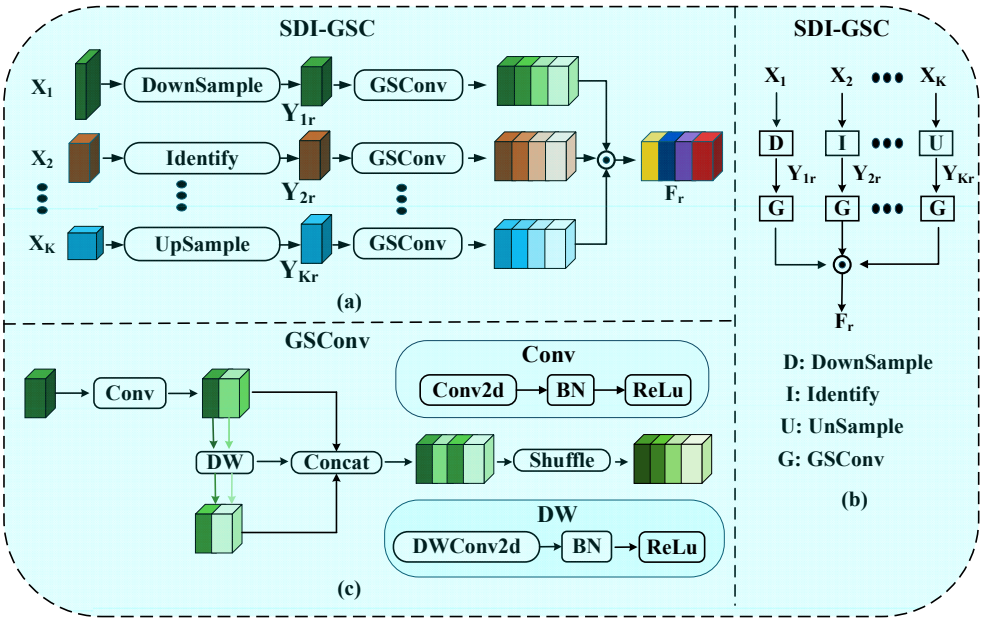


Figure 4. The structure of SDI-GSC. (**a**) SDI-GSC. (**b**) Simplified SDI-GSC. (**c**) GSConv.

Given the input Y_{kr}, the output of GSC is,

$$F_{kr} = Shuffle\left(Concat\left(Conv(Y_{kr}), DWConv\left(Conv(Y_{kr})\right)\right)\right) \tag{8}$$

where $Conv(\cdot)$ represents standard convolution; $DWConv(\cdot)$ represents depth-wise convolution; $Shuffle(\cdot)$ represents the channel shuffling operation.

Through the joint operation of SDI and GSC, the fusion block achieves scale alignment across heterogeneous feature resolutions, semantic-detail complementarity between deep and shallow representations, and efficient channel interaction under lightweight constraints.

Finally, the enhanced feature maps of all *S* branches are fused by the Hadamard multiplication, and the final *r*-th scale output feature maps can be expressed by,

$$F_r = F_{1r} \odot F_{2r} \odot \cdots \odot F_{Kr} \tag{9}$$

where $\odot$ denotes the Hadamard product.

Compared with additive fusion, multiplicative interaction reinforces mutually consistent activations across scales while suppressing inconsistent responses, thereby mitigating clutter-induced false alarms in SAR imagery.

Despite sharing the same SDI-GSC fusion principle, the two structural modes (Mode S and Mode D) adopt different fusion pathways. Under Mode S, shallow feature interaction is emphasized to preserve fine-grained details and reduce over-smoothing. Under Mode D, deeper cross-scale

 doi:10.20944/preprints202605.0337.v1

aggregation paths are activated to enhance contextual reasoning and suppress clutter interference. The detailed structures of these two modes are shown in Figure 5(a) and (b). In Figure 5(a), Mode S adopts SDI-GSC1/2/3 configurations for relatively shallow fusion, while Mode D adopts SDI-GSC4/5/6 configurations for deeper and denser cross-scale interaction. The simplified structures of configurations are shown in Figure 5(c)-(h), respectively.

3.5. CIoU-Modulated Head

YOLOv8 adopts an anchor-free detection paradigm, in which each spatial prediction location on the feature map directly outputs classification scores and bounding box regression parameters. In its detection head, a task-aligned learning (TAL) mechanism is employed to couple classification confidence with localization accuracy. Specifically, an alignment metric—typically formulated as the product of classification probability and complete intersection over union (CIoU) —is used to jointly represent category confidence and regression quality. However, in SAR target detection, targets such as aircraft and ships are frequently embedded in strong and heterogeneous background clutter. This clutter seriously perturbs the CIoU calculation of predicted bounding boxes, leading to compressed CIoU distributions and reduced separability between high-quality and low-quality predictions. As a result, the alignment metric becomes less discriminative, weakening the reliability of positive sample selection.

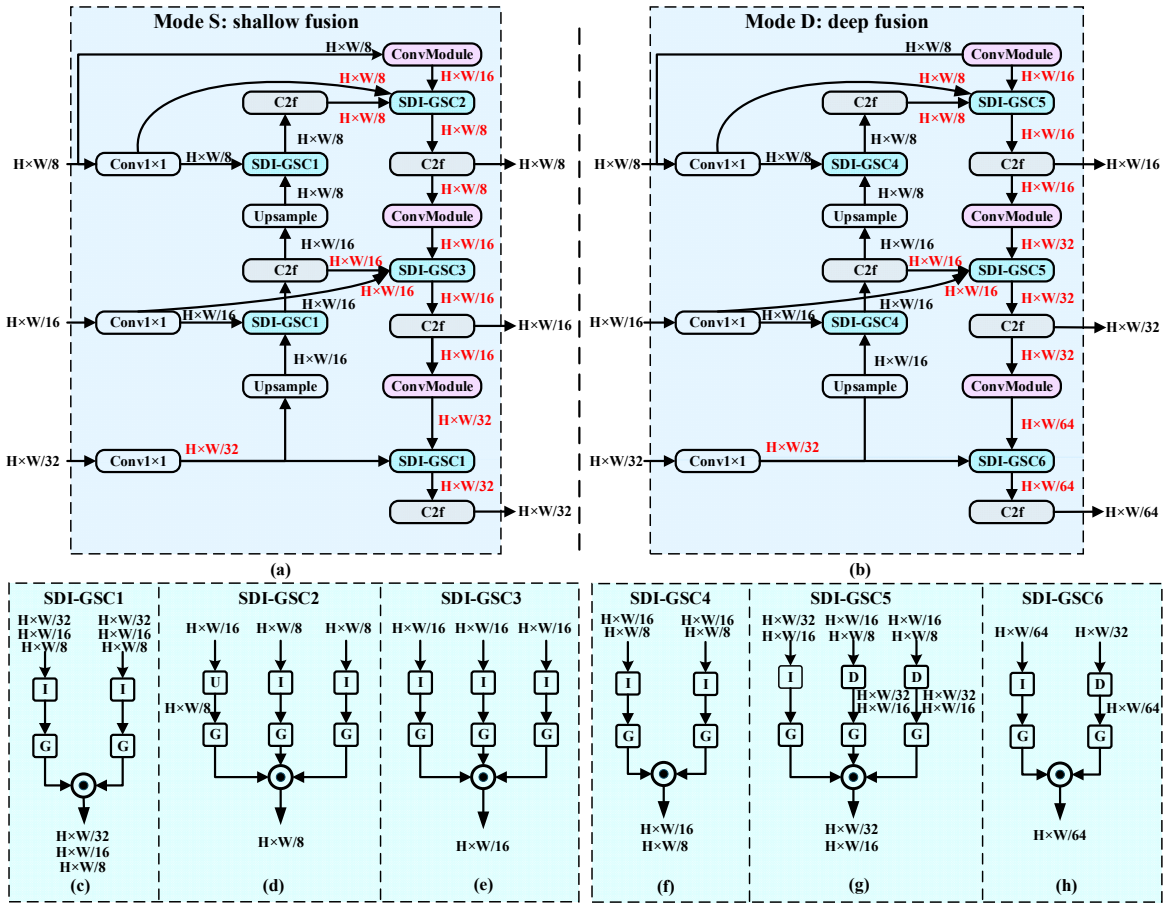


Figure 5. The structures of SAN under Mode S and D. (**a**) Mode S. (**b**) Mode D. (**c**) SDI-GSC1. (**d**) SDI-GSC2. (**e**) SDI-GSC3. (**f**) SDI-GSC4. (**g**) SDI-GSC5. (**h**) SDI-GSC6.

To mitigate this issue, a CMH is proposed, as illustrated in Figure 2. The core idea is to reshape the CIoU distribution before task-aligned computation by amplifying reliable predictions and suppressing clutter-induced unreliable ones. By incorporating this CIoU modulation into the TAL module, the dynamic range of localization quality is enlarged, thereby improving the robustness of

 doi:10.20944/preprints202605.0337.v1

alignment-based sample assignment. Let the batch size be B, the number of ground-truth targets in each image be J, and the total number of prediction locations be N. For the n-th prediction location (n=1, 2, …, N) in the b-th (b=1, 2, …, B) image, the classification head outputs a raw score S_{bn}^{raw}, and the regression head predicts location distributions D_{bn}^{pred}. After applying the sigmoid function, the classification probability is obtained as S_{bn}^{prob}. The regression outputs are decoded into predicted bounding boxes B_{bn}^{pred}.

For the j-th (j=1, 2, …, J) ground-truth box B_{bj}^{gt} in the b-th image, the CIoU between B_{bn}^{pred} and B_{bj}^{gt} is computed as,

$$I_{bjn} = \mathrm{CIoU}\left(B_{bn}^{Pred}, B_{bj}^{gt}\right) \tag{10}$$

During the task alignment, the CIoU is nonlinearly modulated to expand the separability between reliable and unreliable predictions,

$$I'_{bjn} = \begin{cases} \left(I_{bjn}\right)^{1/Power} & I_{bjn} \geq I_0 \\ \left(I_{bjn}\right)^{Power^2} & I_{bjn} < I_0 \end{cases} \tag{11}$$

where I_0 represents the threshold, and $Power = 2$.

This modulation increases the contrast between high-CIoU and low-CIoU samples. High-quality predictions receive amplified supervision signals, whereas clutter-induced low-overlap predictions are suppressed. From an optimization perspective, this reshaping redistributes gradient contributions toward reliable candidates.

The probability that the n-th prediction location belongs to the category of the j-th ground truth target is defined as,

$$S_{bjn} = S_{bnL_{bj}}^{prob} \tag{12}$$

where L_{bj} represents the category label of the j-th ground-truth box in the b-th image.

The alignment metric is computed by jointly considering classification probability and modulated CIoU,

$$A_{bjn} = \left(S_{bjn}\right)^{\alpha} \times \left(I'_{bjn}\right)^{\beta} \tag{13}$$

where $\alpha = 1$ and $\beta = 6$.

For each ground-truth target, only prediction locations with high alignment scores are retained as candidate positive samples, thereby improving computational efficiency while restricting supervision to geometrically plausible locations. Specifically, for the j-th ground-truth target, select the top K predictions ranked by A_{bjn} as candidate samples by using the mask M_{bjn}^{topK}. Additionally, spatial constraints are enforced so that only predictions located inside the ground-truth box are retained via mask M_{bjn}^{in}.

The positive sample mask of the j-th ground-truth target is defined as,

$$M_{bjn}^{pos,initial} = M_{bjn}^{topK} \times M_{bjn}^{in}. \tag{14}$$

To avoid assigning multiple targets to the same prediction location, conflict resolution is performed. For the n-th prediction location, if it is assigned to multiple ground-truth targets, only the one with the maximum CIoU is retained,

$$M_{bjn}^{pos} = \begin{cases} 1 & if\ j = \underset{j \in J_{bn}^{initial}}{\mathrm{argmax}}\, I_{bjn} \\ 0 & otherwise \end{cases} \tag{15}$$

 doi:10.20944/preprints202605.0337.v1

where $J_{bn}^{initial}=\left\{j\middle|M_{bjn}^{pos,initial}=1\right\}$ represents an index set of ground-truth targets assigned to the n-th prediction location in the b-th image.

The alignment metric is then filtered using the final mask M_{bjn}^{pos},

$$A_{bjn}^{filtered}=A_{bjn}\times M_{bjn}^{pos}. \tag{16}$$

To stabilize training by preventing domination from extreme alignment values, the relative quality score of the n-th prediction location for the j-th ground-truth target is defined as,

$$Q_{bjn}=\frac{A_{bjn}^{filtered}\times I_{bj}^{\max}}{A_{bj}^{\max}} \tag{17}$$

where $A_{bj}^{\max}$ and $I_{bj}^{\max}$ can be calculated by,

$$\begin{cases} A_{bj}^{\max}=\max\limits_{n}A_{bjn}^{filtered} \\ I_{bj}^{\max}=\max\limits_{n}\left(I'_{bjn}\times M_{bjn}^{pos}\right) \end{cases}. \tag{18}$$

For each prediction location n in the b-th image, the final quality weight is determined by taking the maximum relative quality score across all ground-truth targets assigned to the location,

$$W_{bn}=\max_{j}Q_{bjn}. \tag{19}$$

3.6. Loss Function

In the anchor-free YOLOv8 framework, each prediction location outputs classification scores and bounding box regression distributions. Therefore, ground-truth labels must be reformulated into target representations aligned with prediction locations.

For each prediction location n in the b-th image selected as positive samples after task-aligned selection, the corresponding target category label and bounding box are assigned as,

$$\begin{cases} L_{bn}^{target}=L_{b,J_{bn}}^{gt} \\ B_{bn}^{target}=B_{b,J_{bn}}^{gt} \end{cases} \tag{20}$$

where $J_{bn}=\left\{j\middle|M_{bjn}^{pos}=1\right\}$.

Classification Loss: Instead of using a hard binary label, the soft-weighting mechanism allows high-quality positive samples to contribute more to optimization. The soft category label is obtained by weighting the target category label with the final quality weight from task alignment, which can be expressed as,

$$L_{bn}=L_{bn}^{target}\times W_{bn}. \tag{21}$$

With this soft label, the classification loss is formulated using binary cross-entropy:

$$\mathcal{L}_{cls}=\frac{-\sum_{b=1}^{B}\sum_{n=1}^{N}\left[L_{bn}\log\left(S_{bn}^{pred}\right)+\left(1-L_{bn}\right)\log\left(1-S_{bn}^{pred}\right)\right]}{\sum_{b=1}^{B}\sum_{n=1}^{N}W_{bn}}. \tag{22}$$

Regression Loss: Let $fg=\left\{(b,n)\middle|\exists j\,s.t.M_{bjn}^{pos}=1\right\}$ denote the set of prediction locations assigned to at least one ground-truth box. The regression loss can be written as,

$$\mathcal{L}_{reg}=\frac{\sum_{(b,n)\in fg}W_{bn}\cdot\left[1-\mathrm{CIoU}\left(B_{bn}^{\mathrm{pred}},B_{bn}^{\mathrm{target}}\right)\right]}{\sum_{b=1}^{B}\sum_{n=1}^{N}W_{bn}}. \tag{23}$$

Distribution Focal Loss (DFL): YOLOv8 adopts DFL to model bounding box offsets as discrete probability distributions. For the l-th coordinate at prediction location n (where

 doi:10.20944/preprints202605.0337.v1

$l \in \{left, top, right, bottom\}$ indexes the box edge), the predicted distribution d_t^{bnl} is converted to a probability distribution via softmax,

$$p_t^{bnl} = \frac{\exp\left(d_t^{bnl}\right)}{\sum_{t'=0}^{Regmax-1} \exp\left(d_{t'}^{bnl}\right)}, \quad t = 0, 1, \ldots, Regmax-1 \tag{24}$$

where $Regmax$ is the maximum regression bin index.

For the target distribution y^{bnl} encoded from B_{bn}^{target}, the left and right bin indices and the corresponding interpolation weights are defined as,

$$\begin{cases} t_{left} = \lfloor y^{bnl} \rfloor \\ t_{right} = t_{left} + 1 \\ W_{left} = t_{right} - y^{bnl} \\ W_{right} = 1 - W_{left} \end{cases} \tag{25}$$

where $\lfloor \cdot \rfloor$ represents the floor operation.

The DFL loss is computed as,

$$\mathcal{L}_{dfl} = \frac{\sum_{(b,n) \in fg} W_{bn} \left[\sum_{l=1}^{4} \mathcal{L}_{dfl}^{0}\left(p^{bnl}, y^{bnl}\right) \right]}{\sum_{b=1}^{B} \sum_{n=1}^{N} W_{bn}} \tag{26}$$

where

$$\mathcal{L}_{dfl}^{0}\left(p^{bnl}, y^{bnl}\right) = -\left[W_{left} \cdot \log\left(p_{t_{left}}\right) + W_{right} \log\left(p_{t_{right}}\right) \right]. \tag{27}$$

Finally, the total loss can be expressed by,

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{dfl}\mathcal{L}_{dfl} \tag{28}$$

where λ_{cls}, λ_{reg}, and λ_{dfl} are balancing coefficients.

It is worth noting that the proposed CIoU modulation influences both sample assignment and quality weight, thereby implicitly affecting the optimization of classification, regression, and DFL losses, as reflected in (22), (23), and (26), respectively.

4. Experiments

4.1. Datasets

Experiments are conducted on two representative SAR datasets with distinct complexity characteristics, i.e, SAR-Aircraft-1.0 and HRSID.

SAR-AIRcraft-1.0 consists of 4,368 SAR images acquired by the GaoFen-3 satellite [21], containing 16,463 aircraft instances across seven fine-grained categories. There are Boeing 737, Boeing 787, A220, A320/321, A330, ARJ21, and other categories. This dataset is characterized by complex airport environments, dense target distribution, strong structural interference from airport facilities, multi-scale aircraft instances, and fine-grained classification requirements. Such characteristics result in strong background clutter and heterogeneous scattering patterns. The dataset is split into training, validation, and test sets with a 7:1:2 proportion.

HRSID contains 5,604 SAR images collected from three satellites, with 16,951 ship instances [22]. Unlike SAR-Aircraft-1.0, it includes only a single ship category without fine-grained subdivision. Ships are distributed in offshore and nearshore scenes, accounting for 81.6% and 18.4%, respectively. Since most targets are located in offshore areas with relatively homogeneous backgrounds, the overall dataset complexity is relatively low. The dataset is split into training, validation, and test sets with an 8:1:1 ratio.

 doi:10.20944/preprints202605.0337.v1

4.2. Experimental Setup and Evaluation Metrics

All experiments are conducted on an NVIDIA GeForce RTX 4080 GPU with 16 GB of memory. The implementation is based on Python 3.10.14, PyTorch 2.2.2, and CUDA 12.1. During training, the batch sizes for SAR-AIRcraft-1.0 and HRSID are set to 8 and 32, respectively, and the initial learning rate is 0.01.

Model performance is evaluated from three perspectives: detection accuracy, model complexity, and computational efficiency [42]. Detection accuracy is assessed using precision (P), recall (R), and COCO-style metrics, including AP50, AP75, and AP50:95. Model complexity is assessed in terms of the number of parameters and floating-point operations (FLOPs). Computational efficiency is measured by the inference speed, reported in frames per second (FPS), measured on a single GPU.

4.3. Dataset Complexity Assessment

To quantitatively evaluate dataset-level complexity, the proposed CA-CFAR-based metric is applied to both SAR-Aircraft-1.0 and HRSID. A constant false alarm probability of 2% is adopted. For each pixel, the average background power Z and local adaptive threshold T are computed according to (1)-(3). A unified global reference threshold $T_{global} = 6000$ is used as an intensity benchmark. The empirical proportion of pixels whose adaptive thresholds exceed T_{global} is calculated using (4) and serves as the complexity indicator.

As shown in Figure 6(a) and (b), the complexity indicators for SAR-Aircraft-1.0 and HRSID are 22.3% and 8.2%, respectively. The larger proportion observed in SAR-Aircraft-1.0 indicates stronger background heterogeneity and higher spatial variability of local clutter statistics. In contrast, HRSID exhibits a more homogeneous background distribution at the dataset level. One representative image is randomly selected from each dataset. The pixels with $T \geq T_{global}$ are illustrated in Figure 7(a) and (b), respectively. It can be observed that the high-threshold regions in SAR-Aircraft-1.0 are more densely distributed, whereas those in HRSID are relatively sparse. As a result, the deeper Mode D is selected for the aircraft dataset, while the shallower Mode S is selected for the ship dataset.

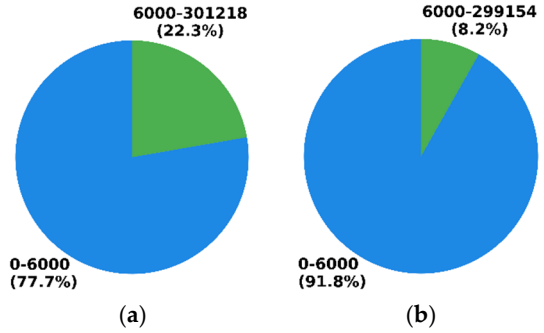


Figure 6. Distribution of local adaptive thresholds. (**a**) SAR-AIRcraft-1.0. (**b**) HRSID.

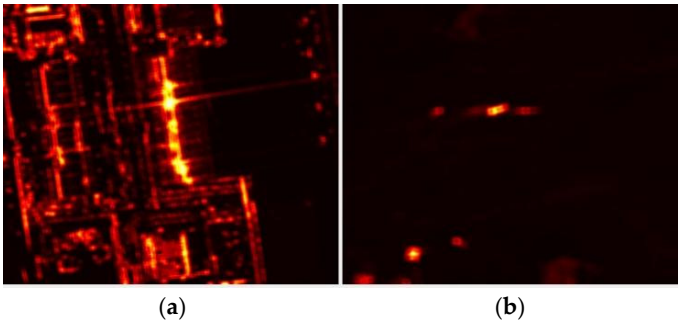


Figure 7. Pixels in an image whose T values exceed T_{global}. (**a**) SAR-AIRcraft-1.0. (**b**) HRSID.

 doi:10.20944/preprints202605.0337.v1

4.4. Ablation Experiment

Ablation experiments are conducted on both SAR-Aircraft-1.0 and HRSID, with the original YOLOv8 serving as the baseline. To ensure fairness and mitigate the risk of overfitting due to limited SAR training data, ResNet18 is used as the backbone in the baseline, while the original FPN-PAN neck and detection head are retained. Based on this baseline, DCSP, SAN, and CMH are progressively incorporated to evaluate their respective effectiveness and complementary contributions.

Results on SAR-Aircraft-1.0 dataset: The ablation results are presented in Table 1. After introducing DCSP into the backbone, P, R, AP50, AP75, and AP50:95 are increased by 0.4%, 0.5%, 0.7%, 0.2%, and 0.4%, respectively. Meanwhile, the number of parameters shows a slight increase, while the inference speed exhibits a marginal decrease, indicating that DCSP enhances feature representation at a low additional cost. Replacing the original FPN-PAN with SAN improves P, R, AP50, AP75, and AP50:95 by 1.3%, 2%, 1.5%, 0.5%, and 0.3%, respectively. The notable gains in precision, recall, and AP50 suggest more efficient multi-scale feature aggregation and better prediction quality in complex airport scenes. In addition, SAN significantly reduces parameters and FLOPs. Although FPS decreases from 285 to 251, this is mainly due to the additional multi-branch fusion and feature alignment operations, which are less hardware-efficient despite their lower theoretical computational cost. Replacing the original detection head with CMH improves P, R, AP50, AP75, and AP50:95 by 2.9%, 2.1%, 2.6%, 0.8%, and 0.7%, respectively. The larger gains in precision, recall, and AP50 indicate that CMH mainly improves overall prediction quality by enhancing the alignment between classification confidence and localization reliability. Consequently, false alarms and missed detections are reduced, while parameters, FLOPs, and FPS remain nearly unchanged.

Table 1. Ablation Experiments on SAR-AIRcraft-1.0 Dataset.

Baseline	DCSP	SAN	CMH	P(%)	R(%)	AP50(%)	AP75(%)	AP50:95(%)	Parameters(M)	FLOPs(G)	FPS
√				85.5	90.5	93.8	80.5	70.6	17.9	46.3	285
√	√			85.9	91	94.5	80.7	71	18.5	46.3	278
√		√		86.8	92.5	95.3	81	70.9	14.3	35.8	251
√			√	88.4	92.6	96.4	81.3	71.3	17.9	46.3	286
√	√	√		87.6	93	95.1	80.9	71.4	14.9	35.8	250
√		√	√	89	92.5	95.8	81.4	71.9	14.3	35.8	253
√	√		√	87.3	94.1	96.6	80.8	71.5	18.5	46.3	279
√	√	√	√	**89.3**	**94.3**	**96.7**	**81.5**	**72**	14.9	35.8	253

For the dual-module settings, DCSP+SAN further improves precision, recall, and AP50:95 over SAN alone, with only slight decreases in AP50 and AP75, indicating a trade-off between stronger feature representation and strict localization accuracy. SAN+CMH achieves clear gains in precision and AP50:95 over SAN alone, demonstrating effective complementarity between feature fusion and reliability-aware prediction modulation. DCSP+CMH yields the most notable gains in recall and AP50 compared with DCSP alone, suggesting improved target coverage and coarse-localization quality. The full model achieves the best performance, with P, R, AP50, AP75, and AP50:95 increased by 3.8%, 3.8%, 2.9%, 1%, and 1.4%, respectively, compared with the baseline. These results demonstrate that DCSP, SAN, and CMH can be effectively integrated to exploit complementary strengths and improve overall detection performance.

Heatmaps for the baseline and the progressively enhanced variants with DCSP, SAN, and CMH are presented in Figure 8(a)-(d), respectively, while the ground truth is given in Figure 8(e). As shown in Figure 8(a), the baseline fails to sufficiently highlight the targets' weak-scattering components, resulting in incomplete activation across the target regions. After introducing DCSP into the backbone, as shown in Figure 8(b), the activations become more concentrated around the dominant scattering centers, and the response intensity over the true target regions is noticeably enhanced. This

 doi:10.20944/preprints202605.0337.v1

improvement can be attributed to the enlarged receptive field and stronger contextual aggregation capability of DCSP. Nevertheless, some parts of the targets still exhibit relatively weak responses. In Figure 8(c), after replacing the original FPN-PAN with SAN, the target regions are well activated, and the overall activation distribution becomes more spatially consistent due to enhanced cross-scale feature interaction and feature alignment. However, a few background regions show weak activations. After further replacing the original detection head with CMH, the heatmap in Figure 8(d) exhibits clearer target focus and sharper response boundaries. In particular, non-target strong scattering regions are significantly suppressed, and the activation responses are better aligned with the ground-truth bounding boxes in Figure 8(e). These observations indicate that the proposed reliability-aware classification-regression alignment effectively mitigates clutter-induced false alarms and improves localization consistency within true target regions.

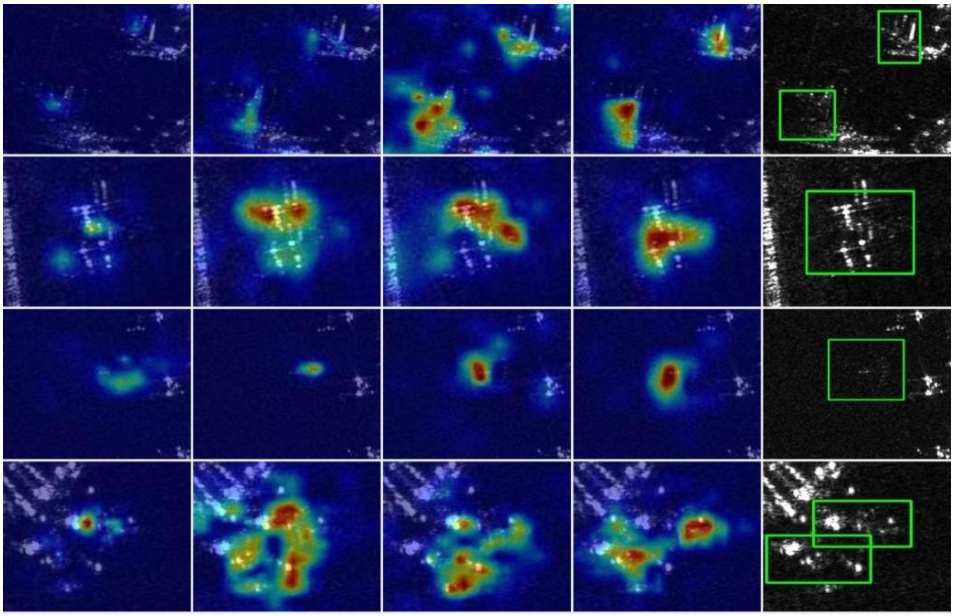


Figure 8. (**a**) Baseline. (**b**) Baseline+DCSP. (**c**) Baseline+DCSP+SAN. (**d**) Baseline+DCSP+SAN+CMH. (**e**) Ground truth.

Results on HRSID dataset: The ablation results are summarized in Table 2. After introducing DCSP into the backbone, the gains in precision and AP50:95 are more pronounced than those observed on the SAR-Aircraft-1.0 dataset. Although recall decreases slightly by 0.1%, this can be regarded as a normal fluctuation. These results suggest that DCSP primarily improves target discrimination rather than recall on this dataset, since ship targets are set against relatively clean backgrounds, making target coverage less challenging. Replacing the original FPN-PAN with SAN yields clear improvements in precision and AP50:95, while recall decreases noticeably. This indicates that SAN effectively suppresses false alarms but tends to produce fewer positive predictions, leading to a more precision-oriented behavior. Replacing the original detection head with CMH yields moderate overall improvements, while AP75 decreases slightly. This indicates that the benefit of CMH in relatively simple scenes is mainly reflected in overall prediction quality rather than further enhancement of high-IoU localization.

For the dual-module settings, DCSP+SAN further improves all evaluation metrics over DCSP alone, demonstrating complementary effects between stronger feature representation and adaptive feature fusion. SAN+CMH improves recall over SAN alone, indicating that CMH can effectively alleviate the conservative prediction tendency introduced by SAN. DCSP+CMH yields a notable improvement in AP75 compared with CMH alone, suggesting that the combination is particularly beneficial for high-quality localization. The complete model yields consistent performance improvements across all evaluation metrics, demonstrating that the three proposed modules

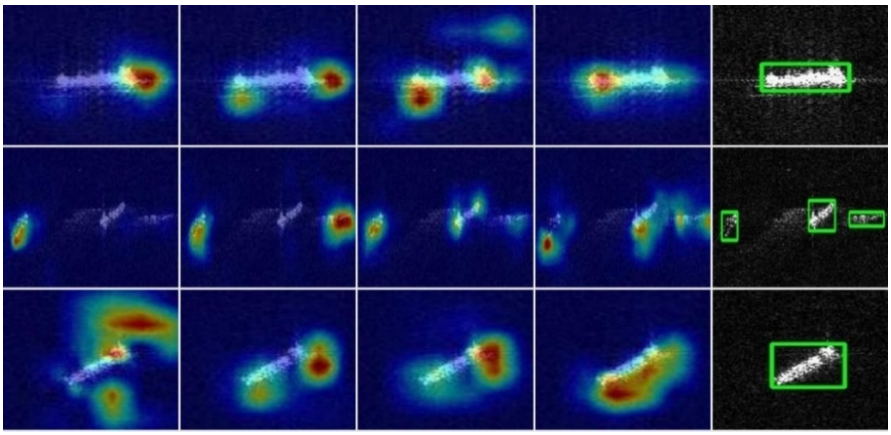
 doi:10.20944/preprints202605.0337.v1

effectively complement one another. In addition, after introducing one or more modules, the variation trends of parameters, FLOPs, and FPS remain generally consistent with those on the SAR-Aircraft-1.0 dataset.

Table 2. Ablation Experiments on HRSID Dataset.

Baseline	DCSP	SAN	CMH	P(%)	R(%)	AP50(%)	AP75(%)	AP50:95(%)	Parameters(M)	FLOPs(G)	FPS
√				90.9	85.7	91.5	79.8	69.4	17.9	46.3	243
√	√			92.2	85.6	92.2	79.9	70.5	18.5	46.3	233
√		√		93.2	84.2	92.2	80	70.4	14.3	43.2	219
√			√	91.6	85.9	92.5	79.7	70.2	17.9	46.3	239
√	√	√		92.9	86.4	93.2	80.7	71.1	14.9	43.2	210
√		√	√	92.5	85.9	92.7	80.6	70.9	14.3	43.2	211
√	√		√	92.6	86.6	91.6	81	70	18.5	46.3	232
√	√	√	√	**93.6**	**87.4**	**93.3**	**81.2**	**71.9**	14.9	43.2	211

Heatmaps for the baseline and for sequentially introducing DCSP, SAN, and CMH to the baseline are presented in Figure 9(a)-(d), respectively; the ground truth is given in Figure 9(e). In Figure 9(a), the phenomena where some target regions fail to be activated, and background clutter is activated, are observed in different image samples. In Figure 9(b), after introducing DCSP, the activation becomes more concentrated around the central target structure. However, some target regions still fail to be activated. In Figure 9(c), after replacing the original FPN-PAN with SAN, all target regions are correctly activated, while some background regions are weakly activated. In Figure 9(d), after further replacing the original detection head with CMH, the target regions are correctly activated, whereas the background clutter is not. Overall, Figure 9 verifies that each proposed component contributes progressively to improved target localization and clutter suppression.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 9. (**a**) Baseline. (**b**) Baseline+DCSP. (**c**) Baseline+DCSP+SAN. (**d**) Baseline+DCSP+SAN+CMH. (**e**) Ground truth.

4.5. Hyperparameter Experiment

The modulation threshold I_0 in (11) determines the boundary between enhancement and suppression for CIoU values in CMH, and therefore plays an important role in reliability-aware task alignment. The detection results under different values of I_0 on SAR-Aircraft-1.0 and HRSID datasets are shown in Tables 3 and 4, respectively.

For the SAR-Aircraft-1.0 dataset with complex background clutter, the model achieves the best performance when I_0 is set to 0.5. By contrast, for the HRSID dataset with relatively simple

 doi:10.20944/preprints202605.0337.v1

backgrounds, the optimal performance is obtained when I_0 is set to 0.1. Except for minor fluctuations in AP75, the other metrics decrease as I_0 increases. This difference in the threshold-parameter experimental results between the two datasets indicates that the optimal modulation threshold is data-dependent. A relatively higher threshold is more suitable for complex and cluttered scenes, where stronger suppression is required to filter unreliable background responses, while a lower threshold is preferable in cleaner scenes to avoid suppressing valid target predictions.

Table 3. Hyperparameter Experiments on SAR-AIRcraft-1.0 Dataset.

I_0	P(%)	R(%)	AP50(%)	AP75(%)	AP50:95(%)
0.1	89.2	92.8	96.3	81.4	71.6
0.3	88.1	94.3	95.9	81.2	71.4
0.5	**89.3**	**94.3**	**96.7**	**81.5**	**72**
0.7	86.4	94	95.1	80.9	70.8

Table 4. Hyperparameter Experiments on HRSID Dataset.

I_0	P(%)	R(%)	AP50(%)	AP75(%)	AP50:95(%)
0.1	**93.6**	**87.4**	**93.3**	**81.2**	**71.9**
0.3	93.1	85.6	92.7	80.8	71.5
0.5	92.6	84.3	91.4	81	71.2
0.7	92.1	82.2	90.3	80.5	70.6

4.6. Comparative Experiment

The effectiveness of the proposed detector is evaluated by comparing it with several state-of-the-art detectors on the SAR-Aircraft-1.0 and HRSID datasets.

Results on the SAR-Aircraft-1.0 dataset: For a fair comparison, all detectors are evaluated using ResNet18 and ResNet50 as backbones, and the results are presented in Table 5. Across both backbone configurations, the proposed CGAAN consistently achieves the best performance among all compared detectors. When using ResNet18, CGAAN outperforms representative anchor-based and anchor-free detectors, including RetinaNet [43], GFL [44], AutoAssign [45], ATSS [46], and FCOS [47]. Compared with more recent advanced detectors, such as RTMDet [48] and YOLOv10 [49], CGAAN consistently improves all evaluation metrics, demonstrating comprehensive performance gains in false-alarm suppression, target coverage, and localization in complex airport scenes. When replacing the backbone with ResNet50, CGAAN still achieves higher recall, AP50, and AP75 than Faster R-CNN [50], Cascade R-CNN [51], RepPoints [52], SKG-Net [53], and SA-Net [21], demonstrating its robustness across different backbone configurations. However, compared with the ResNet18-based setting, the performance decreases. This indicates that a deeper backbone is not more suitable for small-sample SAR datasets.

The detection results of all models using ResNet18 as the backbone are visualized for comparison. Five randomly selected images are used for qualitative comparison, with the results of eight detectors presented in Figure 10(a)-(h). In these figures, green, red, and blue bounding boxes denote correctly detected targets, missed detections, and false alarms, respectively. As shown in Figure 9(a)-(e), early detectors such as RetinaNet, GFL, AutoAssign, ATSS, and FOCS suffer from false alarms and missed detections, and duplicate detections frequently occur. This observation indicates insufficient discrimination between targets and complex background clutter. In Figure 10(f)-(g), corresponding to RTMDet and YOLOv10, the number of missed detections is significantly reduced. However, false alarms remain relatively prominent, suggesting that although these methods improve target coverage, their ability to suppress clutter-induced responses is still limited. In Figure 10(h), the proposed CGAAN further reduces both false alarms and missed detections while

 doi:10.20944/preprints202605.0337.v1

alleviating duplicate detections. Moreover, the predicted bounding boxes exhibit better spatial consistency with the ground-truth targets, indicating improved localization accuracy. These observations are consistent with the metrics achieved by CGAAN in Table 5.

Table 5. Comparative Experiments on SAR-AIRcraft-1.0 Dataset.

Methods	Backbone	P(%)	R(%)	AP50(%)	AP75(%)
RetinaNet [43]	ResNet18	76	71.9	79	56.4
GFL [44]	ResNet18	80.9	79.2	83.9	59.1
AutoAssign [45]	ResNet18	82.3	80	85.5	68.2
ATSS [46]	ResNet18	75.2	74.9	80.3	59.1
FCOS [47]	ResNet18	78.6	79.8	85.6	61
RTMDet [48]	ResNet18	82.6	92.6	94.2	73.9
YOLOv10 [49]	ResNet18	87.1	91.8	95.3	78.9
CGAAN (Ours)	ResNet18	**89.3**	**94.3**	**96.7**	**81.5**
Faster R-CNN [21,50]	ResNet50	77.6	78.1	71.6	53.6
Cascade R-CNN [21,51]	ResNet50	**89**	79.5	77.8	59.1
RepPoints [21,52]	ResNet50	62.7	88.7	80.3	52.9
SKG-Net [21,53]	ResNet50	57.6	88.8	79.8	51
SA-Net [21]	ResNet50	87.5	82.2	80.4	61.4
CGAAN (Ours)	ResNet50	87	**94.1**	**95.6**	**80.6**

Results on the HRSID dataset: The comparison results under the ResNet18 backbone are presented in Table 6. The proposed CGAAN delivers the most favorable performance among all compared detectors. Compared with conventional detectors such as RetinaNet [43], GFL [44], AutoAssign [45], ATSS [46], FCOS [47], DDOD [54], and FoveaBox [55], CGAAN yields consistent improvements across all metrics. It also outperforms recent advanced methods, including RTMDet [48] and YOLOv10 [49], particularly in AP75, indicating better localization quality under stricter evaluation criteria. Although HRSID has a relatively simple background, clear performance differences among detectors still exist. These results suggest the effectiveness and stable performance of the proposed CGAAN in relatively less challenging SAR scenes.

Table 6. Comparative Experiments on HRSID Dataset.

Methods	Backbone	P(%)	R(%)	AP50(%)	AP75(%)
RetinaNet [43]	ResNet18	83.9	69.2	78.8	59.8
GFL [44]	ResNet18	91.1	71.6	82.8	62.1
AutoAssign [45]	ResNet18	88.7	73.7	83	62.7
ATSS [46]	ResNet18	87	71.7	81.8	61.8
FCOS [47]	ResNet18	88.6	70.4	81	61.9
DDOD [54]	ResNet18	83.3	59.7	70.3	57.6
FoveaBox [55]	ResNet18	83.8	65.2	75.5	59
RTMDet [48]	ResNet18	93	82.5	90.5	71.3
YOLOv10 [49]	ResNet18	90.9	83.1	90	73.8
CGAAN (Ours)	ResNet18	**93.6**	**87.4**	**93.3**	**81.2**

The detection results of RetinaNet, GFL, AutoAssign, ATSS, FCOS, DDOD, FoveaBox, RTMDet, YOLOv10, and the proposed CGAAN, all using ResNet18 as the backbone, are visualized in Figure 11. Three offshore and two nearshore images are selected for comparison. From left to right, the 2nd and 5th columns correspond to offshore images, while the 1st, 3rd, and 4th columns correspond to nearshore images. The color definitions are consistent with Figure 10. As shown in Figure 11(a)-(g),

 doi:10.20944/preprints202605.0337.v1

most detectors suffer from false alarms and missed detections. In Figure 11(h)-(i), RTMDet and YOLOv10 significantly reduce missed detections, especially in offshore scenes, while suppressing false alarms to some extent. In Figure 11(j), CGAAN further reduces both missed detections and false alarms across offshore and nearshore scenes. The predicted bounding boxes exhibit better spatial correspondence with the true targets, confirming improved localization consistency. These visual observations are in good agreement with the quantitative results listed in Table 6.

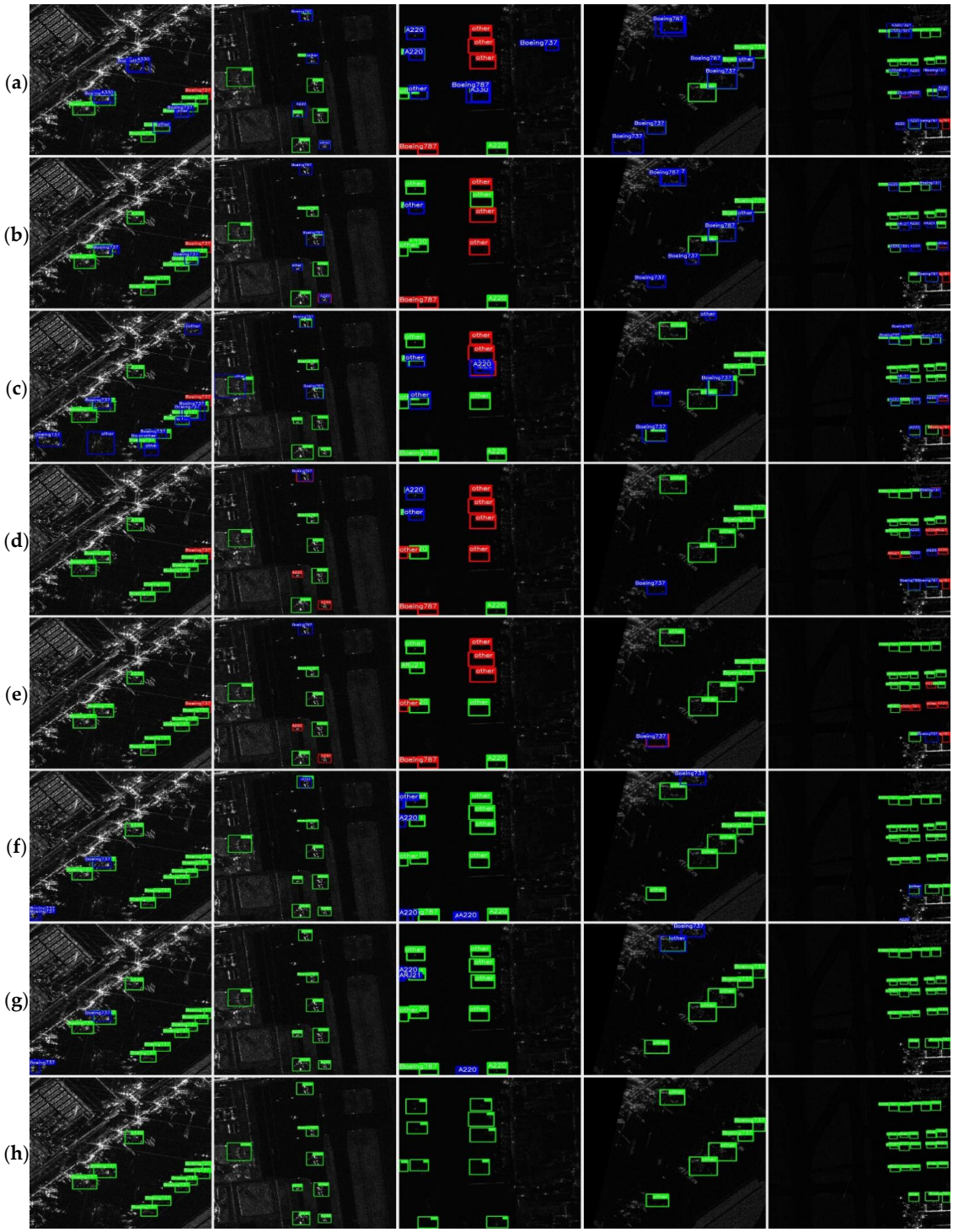


Figure 10. Comparison of detection results on the SAR-AIRcraft-1.0 dataset. (**a**) RetinaNet. (**b**) GFL. (**c**) AutoAssign. (**d**) ATSS. (**e**) FCOS. (**f**) RTMDet. (**g**) YOLOv10. (**h**) CGAAN (Ours). Green, red, and blue bounding boxes denote correctly detected targets, missed detections, and false alarms, respectively.

 doi:10.20944/preprints202605.0337.v1

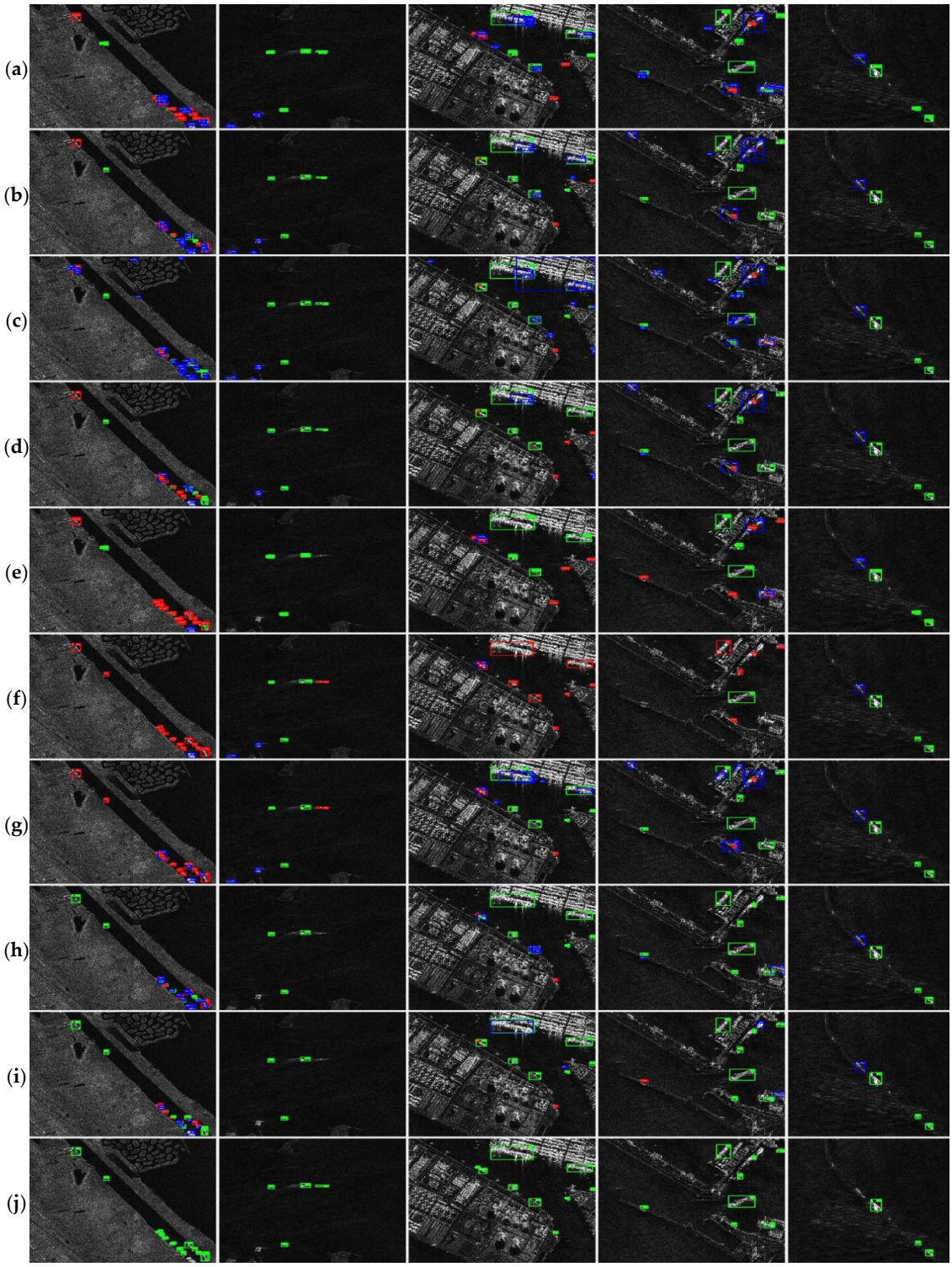


Figure 11. Comparison of detection results on the HRSID dataset. (**a**) RetinaNet. (**b**) GFL. (**c**) AutoAssign. (**d**) ATSS. (**e**) FCOS. (**f**) DDOD. (**g**) FoveaBox. (**h**) RTMDet. (**i**) YOLOv10. (**j**) CGAAN (Ours). From left to right, the 2nd and 5th columns correspond to offshore SAR images, while the 1st, 3rd, and 4th columns correspond to nearshore SAR images.

5. Discussion

5.1. Effectiveness of the Adaptive Structure

Effectiveness of the structure-adaptive fusion mode: To further verify the effectiveness of adapting the fusion model structure to the complexity of the dataset, experiments are conducted with both matched and mismatched feature-fusion configurations. Specifically, the shallow fusion mode

 doi:10.20944/preprints202605.0337.v1

(Mode S) and the deep fusion mode (Mode D) are applied to both datasets. The results under all settings are summarized in Table 7. When the fusion strategy is consistent with the dataset complexity level, significant improvements are observed in P, R, AP50, AP75, and AP50:95. Therefore, the proposed complexity-driven architecture selection mechanism effectively enhances detection performance.

Effectiveness of each module: The ablation results on two datasets show that DCSP, SAN, and CMH all bring performance improvements to different degrees, and their combination further improves the overall detection performance. This demonstrates the complementarity among the three modules. Specifically, DCSP enhances contextual feature representation, SAN improves multi-scale feature fusion, and CMH strengthens the consistency between classification and localization. Meanwhile, the performance gains are achieved without a clear increase in model complexity. Compared with the baseline, the number of parameters and GFLOPs in the full mode is even reduced. Although FPS decreases to some extent, this is mainly caused by hardware-unfriendly operations, rather than an increase in theoretical computational cost.

Effectiveness of the overall detection framework: The comparison results show that the proposed CGAAN outperforms several representative detectors on both SAR-Aircraft-1.0 with complex airport backgrounds and HRSID with relatively simple ship scenes. This demonstrates the overall detection capability of CGAAN under different SAR scene characteristics. In addition, the visualization results show that CGAAN effectively reduces false alarms, missed detections, and duplicate detections, and produces more accurate bounding boxes, further verifying its practical detection effectiveness.

Table 7. Detection Results on SAR-AIRCRAFT-1.0 and HRSID Datasets under Two Fusion Modes.

Dataset	Fusion Mode	P(%)	R(%)	AP50(%)	AP75(%)	AP50:95(%)
SAR-AIRcraft-1.0	D	**89.3**	**94.3**	**96.7**	**81.5**	**72**
	S	85.5	92.2	94.4	79.8	68.2
HRSID	D	89.8	73.2	84.2	73.8	64.2
	S	**93.6**	**87.4**	**93.3**	**81.2**	**71.9**

5.2. Stability of the Adaptive Structure

Stability of performance improvements across datasets: SAR-Aircraft-1.0 contains complex airport backgrounds and fine-grained aircraft targets, whereas HRSID mainly contains ships in relatively simple offshore scenes. Although the two datasets differ significantly in background complexity and scene characteristics, the ablation results show that CGAAN improves P, R, AP50, AP75, and AP50:95 over the baseline on both datasets. This indicates that the proposed detector is not limited to a single target category or scene type, but provides stable performance gains on different datasets.

Stability of complexity and efficiency trends: The number of parameters, GFLOPs, and FPS in the ablation results show generally consistent variation trends on the two datasets. Although the specific GFLOPs and FPS values differ because different fusion depths are adopted by SAN on different datasets, the overall trends remain consistent. This indicates that the proposed structural adaptation is controllable and stable.

Stability of comparison results with different detectors: The comparison results show that CGAAN outperforms representative detectors on both datasets, indicating that its performance advantage is stable rather than being observed only on a single dataset. For SAR-Aircraft-1.0, experiments with different backbone networks further show that CGAAN consistently outperforms other detectors, demonstrating its stability with respect to backbone variations.

 doi:10.20944/preprints202605.0337.v1

6. Conclusions

CGAAN is proposed for SAR target detection across datasets with varying background complexities in this work. First, a CA-CFAR-based complexity metric is introduced to quantify the detection difficulty at the dataset level. Then, guided by the complexity metric, an improved YOLOv8 model is developed to enable adaptive target detection. For the backbone, it is constructed by integrating ResNet18 with a DCSP module, which effectively improves contextual feature modeling while maintaining low computational cost. For the neck, SAN is designed to adjust multi-level feature fusion according to dataset complexity. For the head, CMH is proposed to enhance classification-regression alignment by emphasizing reliable predictions and suppressing clutter-affected ones. Finally, experiments on the SAR-Aircraft-1.0 and HRSID datasets demonstrate the effectiveness and robustness of the proposed CGAAN. Significant improvements are observed on the SAR-Aircraft-1.0 dataset with complex backgrounds, particularly in recall and AP75, indicating enhanced target coverage and localization consistency. For HRSID, where most images exhibit relatively simple backgrounds, CGAAN still achieves stable gains and maintains the best overall performance. The visualization results further confirm that CGAAN can effectively suppress false alarms and missed detections under different scene conditions. These advantages primarily stem from its adaptive structural design. In future work, more flexible adaptive strategies can be explored to further improve generalization across diverse SAR imaging conditions and more challenging small-target detection tasks.

Author Contributions: Conceptualization, L.Y. and W.H.; methodology, X.X.; software, X.X. and X.J.; validation, M.L. and X.Y.; writing, L.Y. and X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (no. 62261027, no. 62561003, no. 62566028, and no. 62266020), the Natural Science Foundation of Jiangxi Province (no. 20252BAC240198, no. 20224BAB202002, and no. 20224BAB212013), the Jiangxi Provincial Graduate Innovation Special Foundation under Grant YC2024-S571, the Jiangxi Provincial Key Laboratory of Multidimensional Intelligent Perception and Control of China (no. 2024SSY03161), and the Supercomputing Platform of Jiangxi University of Science and Technology.

Data Availability Statement: The SAR-Aircraft-1.0 dataset is available at https://aistudio.baidu.com/datasetdetail/312407, accessed on 28 March 2026. The HRSID dataset is available at https://github.com/chaozhong2010/HRSID, accessed on 28 March 2026.

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

References

1. Tian, Y.; Gao, F.; Huang, R.; Wu, Y. HGXES: Lightweight network for ship detection in specific marine environments. *Remote Sens.* 2026, 18, 1276. https://doi.org/10.3390/rs18091276.
2. Ai, J.; Mao Y.; Luo Q; Xing M.; Jiang K.; Jia L. Robust CFAR ship detector based on bilateral-trimmed-statistics of complex ocean scenes in SAR imagery: A closed-form solution. *IEEE Trans. Aerosp. Electron. Syst.* 2021, 57, 1872–1890. https://doi.org/10.1109/TAES.2021.3050654.
3. Li, Y.; Zhang, S.; Wang, W.-Q. A lightweight Faster R-CNN for ship detection in SAR images. *IEEE Geosci. Remote Sens. Lett*. 2022, 19, 4006105. https://doi.org/10.1109/LGRS.2020.3038901.
4. Chai, B.; Nie, X.; Zhou, Q.; Zhou, X. Enhanced cascade R-CNN for multiscale object detection in dense scenes from SAR images. *IEEE Sens. J*. 2024, 24, 20143–20153. https://doi.org/10.1109/JSEN.2024.3393750.
5. Zhou, P.; Niu, B.; Huang, L.; Wang, Q.; Zhao, Y.; Zhou, G.; Hu, Y. SARDet-MIM: Enhancing SAR target detection via a structural and scattering masked autoencoder. *Remote Sens*. 2026, 18, 580. https://doi.org/10.3390/rs18040580.

 doi:10.20944/preprints202605.0337.v1

6. Tan, X.; Leng, X.; Luo, R.; Sun, Z.; Ji, K.; Kuang, G. YOLO-RC: SAR ship detection guided by characteristics of range-compressed domain. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2024, 17, 18834–18851. https://doi.org/10.1109/JSTARS.2024.3478390.
7. Liu, H.; Dong, H.; Shi, H.; Li, F. CCAI-YOLO: A high-precision synthetic aperture radar ship detection model based on YOLOv8n algorithm. *Remote Sens*. 2026, 18, 145. https://doi.org/10.3390/rs18010145.
8. Wang, Z.; Du, L.; Mao, J.; Liu, B.; Yang, D. SAR target detection based on SSD with data augmentation and transfer learning. *IEEE Geosci. Remote Sens. Lett*. 2019, 16, 150–154. https://doi.org/10.1109/LGRS.2018.2867242.
9. Xu, X.; Bai, Y.; Liu, G.; Zhang, P. Lite-YOLOv5: A lightweight deep learning detector for on-board ship detection in large-scene Sentinel-1 SAR images. *Remote Sens*. 2022, 14, 4567. https://doi.org/10.3390/rs14184567.
10. Lin, H.; et al. DCEA: DETR with concentrated deformable attention for end-to-end ship detection in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2024, 17, 17292–17307. https://doi.org/10.1109/JSTARS.2024.3461723.
11. Li, C.; Hei, Y.; Xi, L.; Li, W.; Xiao, Z. GL-DETR: Global-to-local transformers for small ship detection in SAR images. *IEEE Geosci. Remote Sens. Lett*. 2024, 21, 4016805. https://doi.org/10.1109/LGRS.2024.3461212.
12. Ai, J.; Tian, R.; Luo, Q.; Jin, J.; Tang, B. Multi-scale rotation-invariant Haar-like feature integrated CNN-based ship detection algorithm of multiple-target environment in SAR imagery. *IEEE Trans. Geosci. Remote Sens*. 2019, 57, 10070–10087. https://doi.org/10.1109/TGRS.2019.2931308.
13. Ma, X.; Hou, S.; Wang, Y.; Wang, J.; Wang, H. Multiscale and dense ship detection in SAR images based on key-point estimation and attention mechanism. *IEEE Trans. Geosci. Remote Sens*. 2022, 60, 5221111. https://doi.org/10.1109/TGRS.2022.3141407.
14. Yang, S.; An, W.; Li, S.; Wei, G.; Zou, B. An improved FCOS method for ship detection in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2022, 15, 8910–8927. https://doi.org/10.1109/JSTARS.2022.3213583.
15. Yue, T.; Zhang, Y.; Wang, J.; Xu, Y.; Liu, P. A weak supervision learning paradigm for oriented ship detection in SAR image. *IEEE Trans. Geosci. Remote Sens*. 2024, 62, 5207812. https://doi.org/10.1109/TGRS.2024.3375069.
16. Li, Y.; Liu, J.; Li, X.; Zhang, X.; Wu, Z.; Han, B. A lightweight network for ship detection in SAR images based on edge feature aware and fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2025, 18, 3782–3796. https://doi.org/10.1109/JSTARS.2024.3524402.
17. Feng, Y.; Zhang, Y.; Zhang, X.; Wang, Y.; Mei, S. Large convolution kernel network with edge self-attention for oriented SAR ship detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2025, 18, 2867–2879. https://doi.org/10.1109/JSTARS.2024.3514855.
18. Qin, C.; Zhang, L.; Wang, X.; Li, G.; He, Y.; Liu, Y. RDB-DINO: An improved end-to-end transformer with refined de-noising and boxes for small-scale ship detection in SAR images. *IEEE Trans. Geosci. Remote Sens*. 2025, 63, 5200517. https://doi.org/10.1109/TGRS.2024.3515150.
19. Zhou, S.; Zhang M.; Wu L.; Yu D.; Li J.; Fan F. Lightweight SAR ship detection network based on transformer and feature enhancement. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2024, 17, 4845–4858. https://doi.org/10.1109/JSTARS.2024.3362954.
20. Zeng, J.; Tang, X.; Li, S. DAFE-Net: Direction-aware feature enhancement network for SAR ship detection. *Remote Sens.* 2026, 18, 1380. https://doi.org/10.3390/rs18091380.
21. Wang, Z.; Kang, Y.; Zeng, X.; Chen, X.; Zhang, Y.; Fu, K.; et al. SAR-Aircraft-1.0: A high-resolution SAR aircraft detection and recognition dataset. *J. Radars*. 2023, 12, 906–922. https://doi.org/10.12000/JR23043.
22. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access* 2020, 8, 120234–120254. https://doi.org/10.1109/ACCESS.2020.3005861.
23. Luo, C.; Zhang, Y.; Guo, J.; Zhou, G.; You, H.; Li, P.; Ning, X. DEMC: A diffusion-enhanced mutual consistency framework for cross-domain object detection in optical and SAR imagery. *Remote Sens.* 2026, 18, 1358. https://doi.org/10.3390/rs18091358.

 doi:10.20944/preprints202605.0337.v1

24. Zhao, S.; Luo, Y.; Zhang, T.; Guo, W.; Zhang, Z. A feature decomposition-based method for automatic ship detection crossing different satellite SAR images. *IEEE Trans. Geosci. Remote Sens*. 2022, 60, 5234015. https://doi.org/10.1109/TGRS.2022.3201628.
25. Pan, B.; Xu, Z.; Shi, T.; Li, T.; Shi, Z. An imbalanced discriminant alignment approach for domain adaptive SAR ship detection. *IEEE Trans. Geosci. Remote Sens*. 2023, 61, 5108111. https://doi.org/10.1109/TGRS.2023.3303507.
26. Zhang, X.; Zhang S.; Sun Z.; Liu C.; Sun Y.; Ji K. Cross-sensor SAR image target detection based on dynamic feature discrimination and center-aware calibration. *IEEE Trans. Geosci. Remote Sens*. 2025, 63, 5209417. https://doi.org/10.1109/TGRS.2025.3559618.
27. Du, W.; Cheng J.; Zhang C.; Zhao P.; Wan H.; Zhou Z. SARNas: A hardware-aware SAR target detection algorithm via multiobjective neural architecture search. *IEEE Trans. Geosci. Remote Sens*. 2023, 61, 5212923. https://doi.org/10.1109/TGRS.2023.3292618.
28. Kuang, C.; Wang, C.; Wen, B.; Hou, Y.; Lai, Y. An improved CA-CFAR method for ship target detection in strong clutter using UHF radar. *IEEE Signal Process. Lett.* 2020, 27, 1445–1449. https://doi.org/10.1109/LSP.2020.3015682.
29. El-Darymli, K.; McGuire, P.; Power, D.; Moloney, C. Target detection in synthetic aperture radar imagery: A state-of-the-art survey. *J. Appl. Remote Sens*. 2013, 7, 071598. https://doi.org/10.1117/1.JRS.7.071598.
30. Li, T.; Peng, D.; Chen, Z.; Guo, B. Superpixel-level CFAR detector based on truncated gamma distribution for SAR images. *IEEE Geosci. Remote Sens. Lett*. 2021, 18, 1421–1425. https://doi.org/10.1109/LGRS.2020.3003659.
31. Yang, H.; Zhang T.; He Y.; Dan Y.; Yin J.; Ma B. GPU-oriented designs of constant false alarm rate detectors for fast target detection in radar images. *IEEE Trans. Geosci. Remote Sens.* 2022, 60, 5231214. https://doi.org/10.1109/TGRS.2022.3188151.
32. Wang, C.; Guo, B.; Song, J.; He, F.; Li, C. A novel CFAR-based ship detection method using range-compressed data for spaceborne SAR system. *IEEE Trans. Geosci. Remote Sens.* 2024, 62, 5215515. https://doi.org/10.1109/TGRS.2024.3419893.
33. Zhao, Y.; Zhao, L.; Li, C.; Kuang, G. Pyramid attention dilated network for aircraft detection in SAR images. *IEEE Geosci. Remote Sens. Lett*. 2021, 18, 662–666. https://doi.org/10.1109/LGRS.2020.2981255.
34. Guo, Q.; Wang, H.; Xu, F. Scattering enhanced attention pyramid network for aircraft detection in SAR images. *IEEE Trans. Geosci. Remote Sens*. 2021, 59, 7570–7587. https://doi.org/10.1109/TGRS.2020.3027762.
35. Chen, Y.; Cong, Y.; Zhang, L. Deformable scattering feature correlation network for aircraft detection in SAR images. *IEEE Geosci. Remote Sens. Lett*. 2023, 20, 4007205. https://doi.org/10.1109/LGRS.2023.3292243.
36. Kang, Y.; Wang, Z.; Fu, J.; Sun, X.; Fu, K. SFR-Net: Scattering feature relation network for aircraft detection in complex SAR images. *IEEE Trans. Geosci. Remote Sens.* 2022, 60, 5218317. https://doi.org/10.1109/TGRS.2021.3130899.
37. Chen, L.; Luo, R.; Xing, J.; Li, Z.; Yuan, Z.; Cai, X. Geospatial transformer is what you need for aircraft detection in SAR imagery. *IEEE Trans. Geosci. Remote Sens*. 2022, 60, 5225715. https://doi.org/10.1109/TGRS.2022.3162235.
38. Zhou, J.; Xiao, C.; Peng, B.; Liu, Z.; Liu, L.; Liu, Y. DiffDet4SAR: Diffusion-based aircraft target detection network for SAR images. *IEEE Geosci. Remote Sens. Lett*. 2024, 21, 4007905. https://doi.org/10.1109/LGRS.2024.3386020.
39. Luo, R.; He, Q.; Zhao, L.; Zhang, S.; Kuang, G.; Ji, K. Geospatial contextual prior-enabled knowledge reasoning framework for fine-grained aircraft detection in panoramic SAR imagery. *IEEE Trans. Geosci. Remote Sens*. 2024, 62, 5226213. https://doi.org/10.1109/TGRS.2024.3487780.
40. Peng, Y.; Chen, D.Z.; Sonka, M. U-Net V2: Rethinking the skip connections of U-Net for medical image segmentation. In Proceedings of the IEEE 22nd International Symposium on Biomedical Imaging, Houston, TX, USA, 14-17 April 2025.
41. Cui, Z.; Wang, X.; Liu, N.; Cao, Z.; Yang, J. Ship detection in large-scale SAR images via spatial shuffle-group enhance attention. *IEEE Trans. Geosci. Remote Sens*. 2021, 59, 379–391. https://doi.org/10.1109/TGRS.2020.2997200.

 doi:10.20944/preprints202605.0337.v1

42. Chang, H.; Fu, X.; Lang, P.; Guo, K.; Dong, J.; Chang, S. GLDet: Real-time SAR ship detector based on global semantic information enhancement and local gradient information mining. *IEEE Trans. Geosci. Remote Sens*. 2025, 63, 5209020. https://doi.org/10.1109/TGRS.2025.3559551.
43. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020, 42, 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.
44. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang J.; Yang J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6-12 December 2020; pp. 21002–21012.
45. Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; Sun J. AutoAssign: Differentiable label assignment for dense object detection. arXiv 2020, arXiv:2007.03496. https://arxiv.org/abs/2007.03496.
46. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020; pp. 9756–9765.
47. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October - 2 November 2019; pp. 9626–9635.
48. 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. https://arxiv.org/pdf/2212.07784.
49. Ao, W.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding G. YOLOv10: Real-time end-to-end object detection. arXiv 2024, arXiv:2405.14458. https://arxiv.org/pdf/2405.14458.
50. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7-13 December 2015; pp. 1440–1448.
51. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 6154–6162.
52. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. RepPoints: Point set representation for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October - 2 November 2019; pp. 9656–9665.
53. Fu, K.; Fu, J.; Wang, Z.; Sun, X. Scattering-keypoint-guided network for oriented ship detection in high-resolution and large-scale SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2021, 14, 11162–11178. https://doi.org/10.1109/JSTARS.2021.3109469
54. Chen, Z.; Yang, C.; Li, Q.; Zhao, F.; Zha, Z.-J.; Wu, F. Disentangle your dense object detector. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20-24 October 2021; pp. 4939–4948.
55. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. FoveaBox: Beyond anchor-based object detection. *IEEE Trans. Image Process*. 2020, 29, 7389–7398. https://doi.org/10.1109/TIP.2020.30.