remote sensing

MDPI

Article

DA-GDNet: A Data-Augmented Gather-and-Distribute Network for Robust SAR Target Detection

Feihong Zhao [1,2], Yanfeng Li [1,2,*], Wenqian Wu [2], Houjin Chen [2] and Yujing Shang [2]

[1] Tangshan Research Institute, Beijing Jiaotong University, Tangshan 063000, China; 25125215@bjtu.edu.cn
[2] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China; 22120151@bjtu.edu.cn (W.W.); hjchen@bjtu.edu.cn (H.C.); 23120102@bjtu.edu.cn (Y.S.)
* Correspondence: yf.li@bjtu.edu.cn

Highlights

What are the main findings?

- A physics-guided label–extract–screen–compose augmentation pipeline is proposed to generate contextually plausible and detection-oriented SAR training samples.
- DA-GDNet integrates a Gather-and-Distribute mechanism with a Spatial Feature Enhancement Module to improve multi-scale feature fusion and target representation in cluttered SAR scenes.

What are the implications of the main findings?

- The proposed framework improves the robustness of SAR target detection in complex backgrounds and under depression-angle variations in both the MSTAR and ATRNet-STAR datasets.
- The combination of data augmentation and feature enhancement significantly improves detection accuracy and generalization when training data are limited.

Abstract

Synthetic Aperture Radar (SAR) possesses the capacity for all-weather imaging and is widely applied in target detection. However, robust SAR target detection remains challenging due to the limited availability of task-relevant labeled samples that jointly cover target categories, depression angles, and complex target–background contexts. In this paper, we propose a Data-Augmented Gather-and-Distribute Network (DA-GDNet) for SAR image target detection. By jointly optimizing at both the data and architectural levels, the proposed approach enhances the model's capacity for target detection in complex backgrounds. Specifically, we design a SAR image data augmentation strategy that integrates three-dimensional modeling with deep learning. Meanwhile, we incorporate a Gather–Distribute (GD) mechanism and a Spatial Feature Enhancement Module (SFEM) to achieve efficient multi-scale feature fusion and enhance the saliency of target regions. Experimental results on the MSTAR dataset and ATRNet-STAR dataset demonstrate that DA-GDNet not only improves detection accuracy and robustness, but also significantly strengthens the model's adaptability to variations in depression angles and complex backgrounds.

Keywords: data augmentation; gather–distribute mechanism; synthetic aperture radar; target detection

Check for updates

Academic Editors: Prasad S. Thenkabail, Dongdong Wang and Zenghui Zhang

Received: 4 June 2026
Revised: 6 July 2026
Accepted: 14 July 2026
Published: 21 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

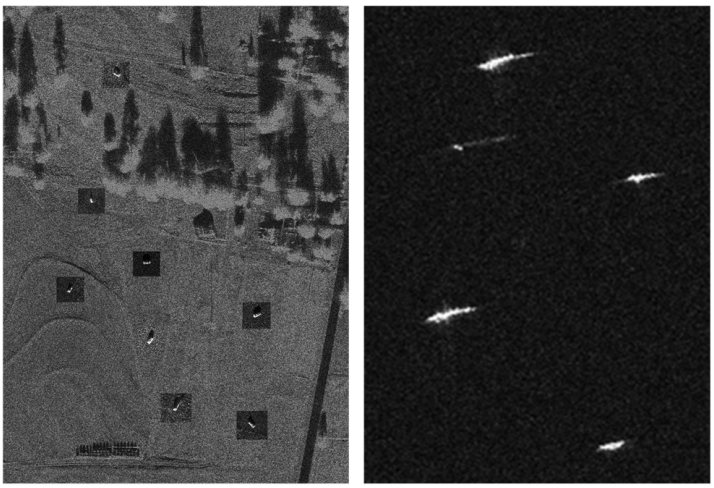
Synthetic Aperture Radar (SAR) is a microwave-based remote sensing technology that offers the advantage of all-weather imaging. Unlike optical remote sensing, whose imaging performance is greatly constrained under adverse meteorological conditions, SAR can penetrate atmospheric disturbances, thereby ensuring stable image acquisition [1]. Due to the distinct scattering characteristics of incident microwaves across different surface objects, SAR exhibits unique strengths in applications such as geological exploration and target recognition. Furthermore, SAR systems are capable of acquiring polarization information through multiple polarization modes. This ability significantly improves object classification and recognition, enabling broad application in resource surveys and environmental monitoring [2].

Spatial resolution is a key factor affecting SAR target interpretation and detection. For small ground or maritime targets, higher spatial resolution can preserve more detailed scattering structures, target contours, and shadow characteristics, which are essential for distinguishing targets from cluttered backgrounds. To obtain high-resolution SAR images, advanced acquisition modes such as Spotlight, Sliding Spotlight, and Staring Spotlight have been widely investigated in spaceborne SAR systems. Compared with conventional stripmap imaging, Spotlight and Sliding Spotlight modes extend the synthetic aperture by steering the antenna beam toward the region of interest, thereby improving azimuth resolution and providing more detailed target observations [3–6]. These high-resolution imaging modes have promoted the development of SAR image interpretation tasks, including object detection, recognition, and fine-grained target analysis. Nevertheless, higher-resolution SAR images also reveal more complex scattering responses, speckle noise, shadows, and background details, which further increase the difficulty of robust target detection in cluttered scenes.

With the rapid development of deep learning [7–9], SAR target detection has achieved remarkable progress in recent years. However, robust detection in complex environments remains challenging. In high-resolution SAR images, small targets often exhibit detailed but unstable scattering responses, and their shadows and surrounding clutter may vary significantly with imaging geometry and depression angle. Therefore, existing methods still suffer from limited generalization under joint variations in target category, depression angle, and background clutter [10–12]. These limitations are closely related not only to detector design, but also to the characteristics of the training data.

It should be emphasized that the data bottleneck in SAR target detection is no longer simply the absolute number of publicly available images. Although recent datasets have considerably increased the scale of SAR data resources [13–15], the number of effective labeled samples available for a specific detection setting is still insufficient. Many datasets remain limited in terms of depression-angle coverage, target diversity, or complex target–background combinations. Consequently, models trained on such data are prone to overfitting to specific imaging geometries and background patterns, and their performances often degrade under unseen angles or cluttered scenes. Therefore, the core data bottleneck in SAR target detection is not the lack of raw SAR images, but the lack of task-relevant labeled samples that jointly cover target categories, depression angles, and physically plausible target–background configurations.

As illustrated in Figure 1, existing SAR datasets such as MSTAR and OpenSARShip may still contain unreasonable target–background configurations or insufficient depression-angle coverage, which motivates the need for more task-relevant and contextually plausible augmented samples [13–16].



(a) (b)

Figure 1. Examples of images from some of the representative SAR datasets. (**a**) Unreasonable data set. (**b**) Single-depression-angle samples.

To alleviate this problem, conventional data augmentation methods enlarge training sets through image-level perturbations such as translation, flipping, noise simulation, and pose variation [17,18]. More recent SAR-specific methods further explore angle- or pose-controllable generation [19–21], which improves target appearance diversity across viewing conditions. However, for SAR target detection, robustness depends not only on angle-related target appearance variation, but also on whether the synthesized samples preserve physically and semantically reasonable target–background relationships during training [22]. In particular, directly compositing whole target images onto new backgrounds is inadequate, because it cannot explicitly control target reuse, placement plausibility, or scene compatibility [23]. By contrast, target segmentation decouples the target signature from limited original scenes and makes the target region controllable [11] so that it can be recomposed into screened background regions to explicitly enlarge the target–background combinational space. This strategy is not intended for arbitrary copy-and-paste augmentation, but for constructing more detection-oriented training samples under complex SAR environments.

Motivated by these observations, we propose a Data-Augmented Gather–Distribute Network (DA-GDNet) for SAR target detection. The proposed augmentation pipeline is designed as a label–extract–screen–compose workflow. Specifically, physics-guided rendering provides target/shadow priors, segmentation extracts controllable target signatures from real SAR images, background screening identifies semantically suitable placement regions, and target–background fusion constructs detection-oriented samples with more plausible contextual relationships. At the architectural level, we further introduce a Gather–Distribute (GD) mechanism and a Spatial Feature Enhancement Module (SFEM) to enhance cross-scale feature fusion and suppress background interference. The main contributions of this paper are summarized as follows:

(1) We propose a physics-guided SAR data augmentation module that integrates SARBake-based target/shadow labeling with U-Net-based segmentation, background suitability screening, and target–background fusion to generate more realistic and detection-oriented training samples.
(2) We introduce a lightweight SAR target detection network with a GD mechanism and an SFEM module to improve multi-scale feature interaction and strengthen target representations in cluttered backgrounds.

(3) Extensive experiments on the MSTAR and ATRNet-STAR datasets demonstrate that the proposed DA-GDNet achieves superior detection accuracy and robustness under varying depression angles and complex backgrounds.

The remainder of this paper is organized as follows. Section 2 reviews related studies on SAR target detection and SAR image data augmentation. Section 3 introduces the materials and methods, including the proposed data augmentation pipeline, the DA-GDNet architecture, datasets, experimental settings, and evaluation metrics. Section 4 presents experimental evaluations on both the MSTAR and ATRNet-STAR datasets, including ablation studies, quantitative comparisons, and qualitative and feature visualization analyses. Section 5 discusses the effectiveness of the proposed data augmentation strategy, the roles of SFEM and GD, and the robustness of DA-GDNet under depression-angle variations and complex clutter. Finally, Section 6 concludes this paper and outlines the limitations and future research directions.

2. Related Work

With the continuous advancement of deep learning technology, SAR image target detection has made remarkable progress. However, due to the intrinsic imaging characteristics of SAR images and the high cost of data acquisition, existing research still faces challenges such as the limited availability of sufficiently diverse and task-relevant labeled samples. In addition, insufficient target feature representation further limits detection performance in complex scenarios. To address these issues, researchers have been actively optimizing detection network architectures to improve model performance, while also extensively exploring data augmentation approaches to mitigate the generalization problems caused by limited data. This section reviews related work in SAR target detection and data augmentation to offer the technical background and theoretical foundation for the proposed approach.

2.1. SAR Target Detection Methods

Traditional SAR target detection methods mainly rely on handcrafted features, statistical modeling, or intensity- and texture-based descriptors. Although these methods are interpretable and effective under relatively simple backgrounds, their detection performance is often limited when target appearances change with imaging geometry, depression angle, and background clutter. With the rapid development of deep learning, SAR target detection has gradually shifted from handcrafted feature design to end-to-end feature representation learning [7,24]. However, compared with optical object detection, robust SAR target detection still faces two major challenges. First, labeled SAR samples that jointly cover target categories, depression angles, and complex target–background configurations are still insufficient. Second, SAR targets are usually small and easily affected by speckle noise, shadows, and cluttered backgrounds, which makes robust target feature extraction more difficult.

2.1.1. Low-Data Learning for SAR Target Detection

To reduce the dependence on large-scale target-level annotations, low-data learning methods, including semi-supervised learning, few-shot learning, and cross-domain transfer learning, have been widely investigated. Liao et al. [25] proposed an improved Faster R-CNN-based semi-supervised SAR target detection framework that combines a decoding module with domain adaptation, allowing unlabeled SAR images and labeled optical remote sensing images to improve detection performance when labeled SAR samples are limited. Shen et al. [26] introduced a threshold filtering semi-supervised learning method for SAR target recognition to alleviate the limited-label problem. Liao [27] further

developed a semi-supervised SAR target detection framework based on YOLOv5 and cross-domain transfer learning, aiming to transfer useful knowledge from optical imagery to SAR data. Similarly, Luo et al. [28] proposed SAR-CDSS, a semi-supervised cross-domain object detection framework from optical to SAR domain, which reduces domain shifts at the image, instance, and feature levels. These methods effectively reduce annotation costs and improve model learning when labeled SAR samples are scarce. Nevertheless, they mainly optimize the learning strategy or transfer process, while the diversity of physically plausible target–background combinations is not explicitly expanded. As a result, their robustness may still degrade under unseen depression angles, complex clutter, or new target-scene configurations.

2.1.2. Few-Shot SAR Target Detection

Few-shot SAR target detection methods attempt to recognize novel categories using only a small number of annotated examples. Zhou et al. [29] proposed Gaussian meta-feature balanced aggregation for few-shot SAR target detection, which models feature distributions to reduce the bias toward base classes. Wang et al. [30] designed fine-grained prototype distillation to improve the representation of novel categories by distilling discriminative support features into detailed prototypes. These prototype- or meta-learning-based methods are effective for improving the adaptability of detectors under extremely limited samples. However, few-shot detectors mainly focus on feature transfer from base classes to novel classes, and their performance still depends strongly on whether the available samples can adequately represent target variations in depression angles, scattering characteristics, and background contexts. Therefore, few-shot learning alone is insufficient to solve the data diversity problem in complex SAR detection scenarios.

2.1.3. Attention Mechanisms and Lightweight Feature Enhancement

In addition to low-data learning, many studies improve SAR target detection by enhancing feature representation in network architectures. Since SAR targets are often small and easily confused with clutter, attention mechanisms and lightweight feature enhancement modules have been introduced to highlight target regions and suppress irrelevant background responses. SENet and ECA-Net enhance important feature responses through channel recalibration [31,32], whereas CBAM and Coordinate Attention further integrate spatial position and channel dependencies to improve the model's focus on target regions [33,34]. In SAR detection, SAR-ShipNet combines bidirectional coordinate attention with multi-resolution feature fusion to enhance ship-target responses [35]. AMANet aggregates features from different levels through an adaptive multi-hierarchical attention module, thereby improving small-target detection in complex coastal backgrounds [36]. RSNet further explores lightweight architectural design to balance parameter efficiency and detection accuracy [37]. These attention-based or lightweight feature-enhancement methods can improve the discriminability of target features under cluttered backgrounds. However, most of them focus on enhancing features after they have already been extracted and may not fully address the joint influence of local scattering details, global spatial context, and depression-angle-induced target deformation.

2.1.4. Multi-Scale Feature Fusion

Multi-scale feature fusion is another important research line for SAR target detection, especially for small targets. FPN combines deep semantic information with shallow spatial details through a top-down pathway and lateral connections [38]. PANet further strengthens bottom-up path aggregation [39], while BiFPN introduces learnable weights to achieve more efficient multi-scale feature fusion [40]. In SAR detection, Swin-FEFPN improves SAR ship detection in complex backgrounds by combining Swin Transformer with a

feature-enhancement pyramid structure [41]. CASS-Det enhances multi-scale dense target detection through center enhancement, neighbor attention, and a cross-connected FPN [42]. These multi-scale feature pyramid methods improve the ability to detect targets of different sizes by fusing features from different network stages. However, most FPN-like structures rely on layer-by-layer feature propagation. During repeated upsampling, downsampling, and convolution operations, fine-grained scattering textures and localization cues of small SAR targets may be gradually weakened. This information attenuation becomes more severe in cluttered scenes and under large depression-angle variations.

Overall, existing SAR target detection methods have improved detection robustness from different perspectives. Semi-supervised, few-shot, and transfer learning methods reduce the dependence on large-scale labeled SAR samples, but they do not explicitly increase the diversity of physically plausible target–background configurations. Attention-based and lightweight detectors enhance target saliency and suppress clutter responses to some extent, but they are usually insufficient to jointly model local target details, global contextual information, and depression-angle variations. Multi-scale feature fusion methods improve small-target representation, but conventional feature pyramid structures may still suffer from cross-layer information attenuation. Therefore, robust SAR target detection under limited labeled samples, complex clutter, and multi-angle variations still requires a unified framework that jointly improves data diversity and target feature representation. Motivated by this gap, the proposed DA-GDNet combines a physics-guided data augmentation strategy with the Spatial Feature Enhancement Module (SFEM) and Gather-and-Distribute (GD) mechanism, aiming to generate more detection-oriented SAR samples and enhance multi-scale target representations in cluttered scenes.

2.2. SAR Image Data Augmentation Methods

In SAR target detection tasks, data augmentation plays an important role in alleviating the shortage of task-relevant labeled samples and improving model generalization under complex imaging conditions. Existing SAR data augmentation methods can be broadly divided into three categories: image-level perturbation augmentation, generative augmentation, and composition-based augmentation.

2.2.1. Image-Level Perturbation Augmentation

Early SAR data augmentation studies mainly relied on image-level transformations to enlarge the training set. Typical operations include random cropping, translation, flipping, speckle noise simulation, and pose synthesis. Furukawa et al. [24] investigated random crop-based augmentation to improve the translation invariance of CNN-based SAR target classification models. Ding et al. [18] further combined translation perturbation, speckle noise simulation, and pose synthesis to improve the robustness of CNN models under limited SAR samples. In addition, Ding et al. [43] proposed a multilevel reconstruction strategy based on attributed scattering centers to generate augmented SAR target samples from a scattering-feature perspective. These methods are simple and effective for increasing local appearance diversity. However, they mainly operate on target chips or image-level perturbations and therefore have limited ability to construct new target–background combinations or simulate complex contextual variations in detection scenarios.

2.2.2. Generative Augmentation

To overcome the limited diversity of handcrafted perturbations, generative models have been widely explored for SAR image synthesis. Cui et al. [44] introduced a GAN-based SAR image augmentation method to expand SAR sensor image samples. Xie et al. [45] employed InfoGAN to control category and pose-related latent variables for SAR image generation. Wang et al. [20] proposed an azimuth-controllable GAN to synthesize SAR target images with controllable azimuth angles, while Zeng et al. [19] developed ATGAN to generate SAR target images for automatic target recognition with improved angle-related controllability. For detection-oriented generation, Ju et al. [46] proposed a GAN-based SAR image generation method for oriented ship detection, and Ju et al. [22] further developed SARGAN for SAR ship detection by incorporating task-related constraints. These methods effectively improve target appearance diversity and angle-related sample coverage. Nevertheless, most generative approaches focus primarily on synthesizing target appearance or complete image samples, while the physical and semantic compatibility among target scale, placement location, shadow distribution, and surrounding background is still not explicitly guaranteed.

2.2.3. Composition-Based Augmentation

Another line of research attempts to enrich training data by recomposing target and background information. In general computer vision, copy-paste augmentation has shown that instance-level recomposition can improve data efficiency, especially when annotated samples are limited [47]. Context-aware augmentation further demonstrates that object placement should consider whether the surrounding scene is suitable, because random pasting may introduce unrealistic samples and hurt detection performance [23]. In SAR imagery, Geng et al. [48] explored clutter transfer to synthesize SAR samples with diverse background clutter, showing the importance of background variation for SAR target recognition. Zhao et al. [49] incorporated target and shadow information and proposed a geometry-based augmentation strategy to compensate for shadow distortion under different depression angles. These studies indicate that target–background composition and shadow-related geometry are important for SAR data augmentation. However, existing composition-based methods still do not sufficiently address the joint control of target extraction, placement suitability, target–shadow consistency, and background compatibility in complex SAR detection scenes.

To address these limitations, our method introduces a structured label–extract–screen–compose augmentation workflow. Specifically, physics-guided target/shadow labeling provides prior information for target and shadow regions; segmentation-based target extraction decouples target signatures from limited original scenes; background suitability screening identifies semantically reasonable placement regions; and target–background fusion constructs detection-oriented training samples with more plausible contextual relationships. Compared with image-level perturbation and generative augmentation methods, the proposed strategy not only enriches depression-angle-related target appearance, but also explicitly expands the target–background combinational space for robust SAR target detection.

3. Materials and Methods

To enhance SAR target detection accuracy and generalization, we propose a unified framework integrating data-level augmentation with architecture-level optimization. At the data level, a dedicated module synthesizes physically plausible and detection-oriented SAR samples by incorporating target–shadow geometric priors and background suitability constraints. Architecturally, the improved network incorporates a Gather–Distribute mech-

anism for multi-scale feature fusion and a spatial feature enhancement strategy to highlight targets while suppressing background interference. This joint optimization effectively bridges the gap between limited data resources and robust detection in complex scenarios.

3.1. Data Augmentation

To achieve high-quality SAR image data augmentation, our augmentation module integrates three core components: target segmentation, background region classification, and target–background fusion. This design ensures both semantic plausibility and physical interpretability of the synthesized data, while simultaneously providing diverse and realistic samples to improve the generalization capability of the subsequent detection network.

3.1.1. Target Segmentation

The SARBake algorithm [50] is a synthetic label-generation method based on high-precision physical modeling that can achieve precise annotation of target and background regions in SAR images through the combined use of 3D modeling and geometric projection.

The role of SARBake in our framework is not to fully simulate the electromagnetic scattering process, but to provide geometry-based target and shadow priors derived from SAR imaging geometry. In SAR images, the target body and its radar shadow are closely related to the radar line-of-sight direction, target geometry, and depression angle. Compared with manually cropped target patches or purely image-level transformations, SARBake-derived labels provide explicit target/shadow separation and thus offer more reliable geometric supervision for subsequent target and shadow extraction.

This method approximates the geometric projection process of SAR imaging using a 3D CAD model as the basis. In the rendering stage, SARBake independently renders the target body and the ground region to construct a complete imaging scene. Subsequently, a geometric projection method is applied to analyze each radar line of sight. It calculates the distance between surface points on the target and their corresponding points on the ground to identify and determine regions with potential occlusion relationships. Based on this analysis, the shadow regions cast by the target are automatically generated and distinguished from both the target and the background, resulting in precise pixel-level label maps comprising target, shadow, and background. The detailed procedure of this algorithm is shown in Algorithm 1.

Although SARBake provides geometry-based target/shadow labels from 3D CAD models, pure geometric projection cannot fully account for speckle noise, complex background clutter, and scattering variations in real SAR images. To bridge the gap between geometry-based priors and real SAR pixel distributions, we employ U-Net [51] as a supervised segmentation network trained with SARBake-derived labels. Specifically, SARBake first generates pixel-level label maps that distinguish the object, its radar shadow, and the background according to the imaging geometry. These label maps are then used to supervise U-Net training on real SAR images. After training, U-Net predicts the object and shadow masks, which are used to extract the corresponding regions for subsequent data augmentation. This integration improves target-extraction accuracy and supports better generalization across depression angles, providing more reliable training samples for subsequent target detection.

Algorithm 1 SARBake Algorithm

Input: Target model M, ground plane G_p, viewpoint set $R_{dir} = \{r_1, r_2, \ldots, r_n\}$, rendering resolution $W \times H$, camera parameters (radius R_c, viewport size S_c, near and far clipping planes Z_n, Z_f), and viewpoint file F_v.

Output: Depth map set $D = \{D_1, D_2, \ldots, D_N\}$ and pixel label map L_{map}.

Step 1: Scene and Parameter Initialization.

Initialize the rendering environment and parse the input parameters. Load the target model M and perform unit normalization. Construct the ground plane G_p as a reference geometry. Extract the predefined viewpoint set R_{dir} from the viewpoint file F_v, where each viewpoint is defined by an elevation angle θ_i and an azimuth angle φ_i.

Step 2: Separate Rendering of Target and Ground.

Use an orthographic projection camera to render the target model M and the ground plane G_p separately.

Step 3: Far-Field Viewpoint Setup and Depth Map Generation.

For each viewpoint $r_i \in R_{dir}$:

1. Compute the camera position: $c_i = (R_c \cdot \sin\theta i \cdot \cos\varphi i, R_c \cdot \sin\theta i \cdot \sin\varphi i, R_c \cdot \cos\theta i)$.

2. Set the camera orientation toward the origin $o = (0, 0, 0)$, with the viewport range defined as $[-S_c/2, S_c/2] \times [-S_c/2, S_c/2]$.

3. Render the depth buffer to generate the depth map D_i.

Step 4: Definition of Radar Line-of-Sight Directions.

Define the viewpoint set R_{dir} as the collection of radar line-of-sight directions, where each line-of-sight r_i corresponds to a depth map D_i.

Step 5: Depth Map Processing and Occlusion Analysis.

For each line-of-sight r_i and its corresponding depth map D_i:

1. Extract the set of depth values $P_i = \{p_1, p_2, \ldots, p_N\}$ and sort them in ascending order by distance, where p_j denotes the depth value of a pixel.

2. Initialize the occlusion reference point as $p_0 = p_1$.

3. For each pixel point p_j ($j = 2, \ldots, N$), compute the theoretical height $h_j{}^t$ at position p_j along the occlusion line extended from p_0.

4. If $p_j < h_j{}^t$, mark p_j as a shadow pixel; otherwise, mark p_j as a target pixel or ground pixel depending on the rendered object, and update p_0.

Step 6: Coordinate System Mapping.

Map each point in the rendered image to the radar coordinate system (x, r), where x denotes the horizontal coordinate on the image plane and r represents the depth value p_j from the camera to the point.

Step 7: Depth Map Resampling and Label Map Generation.

For each depth map D_i, perform column-wise resampling to obtain the corresponding pixel label map $L_{map,i}$.

Combine all $L_{map,i}$ to construct the global label map L_{map}.

Step 8: Output Results.

Return the set of depth maps D and the label map L_{map}.

3.1.2. Background Region Classification

In synthetic data generation, arbitrarily placing target patches on background images may introduce semantic inconsistencies and physically implausible samples, such as vehicles appearing in lakes, dense forests, or overlapping with existing structures. Such unrealistic composites can confuse the detector and degrade its generalization. To avoid this problem, we formulate background suitability estimation as a patch-level binary classification task and employ ResNet18 [52] to distinguish placeable from non-placeable SAR background regions. ResNet18 is adopted because it provides sufficient representation

capability with a lightweight structure, reducing the risk of overfitting under limited manually annotated background patches.

Specifically, to train the model, ten background images with size of a 1400 × 1700 pixels are randomly selected. Using a non-overlapping sliding window strategy, each image is divided into multiple non-overlapping regions of size 100 × 100. These candidate regions are then manually annotated as either placeable or non-placeable. The non-placeable regions primarily correspond to areas where targets are already present or where background features (such as forests or man-made structures) make target placement unsuitable, as illustrated in Figure 2.

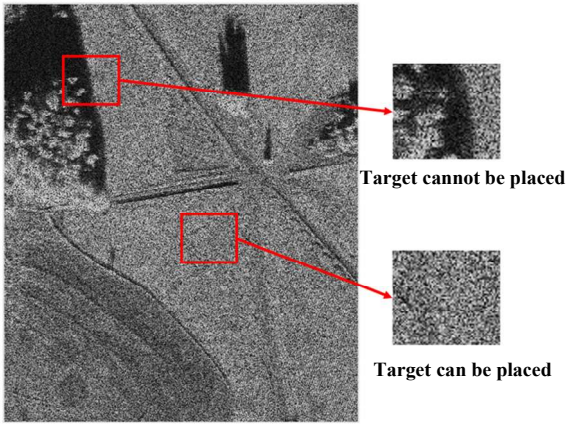


Figure 2. Schematic diagram of background region classification method.

After annotation, these patches and corresponding labels are used as supervised data for training the ResNet18 network. During training, the network employs the cross-entropy loss to quantify the error between the predicted classes and the ground truth labels. During the testing phase, the SAR images in the test set are similarly divided into 100 non-overlapping candidate regions of size 100 × 100 using the sliding window approach, and the trained ResNet18 model is then used to perform classification predictions on these regions.

Finally, a binary mask with the same dimensions as the original SAR image is generated. Through the above background region classification method, the accuracy and rationality of subsequent target detection and placement are ensured.

It should be noted that the background classifier serves solely as a data-quality control component in the offline academic augmentation pipeline. Its function is limited to reducing visually implausible synthetic composites and does not involve the evaluation of tactical suitability, geographic importance, accessibility, or operational deployment conditions.

3.1.3. Target–Background Fusion

If we directly put the target patch on the background image, noticeable artificial boundaries may appear. To ensure natural and coherent integration between the target object and the background image, the target object is first preprocessed through operations such as rotation, flipping, and normalization. These steps increase data diversity while maintaining consistency. Subsequently, 50 candidate regions of size 100 × 100 pixels are randomly selected using a non-overlapping sliding window strategy. The trained background classification network is then employed to evaluate these candidate regions, from which 5 to 15 regions are randomly chosen from the top 20 with the highest confidence scores for target placement. This strategy balances placement reliability and sample diversity: selecting from top-ranked regions avoids physically or semantically unreasonable placement, while random sampling prevents repeated placement in visually similar backgrounds. The

number of selected regions is limited to 5–15 to avoid excessive target density in a single SAR scene. The overall process is illustrated in Figure 3.

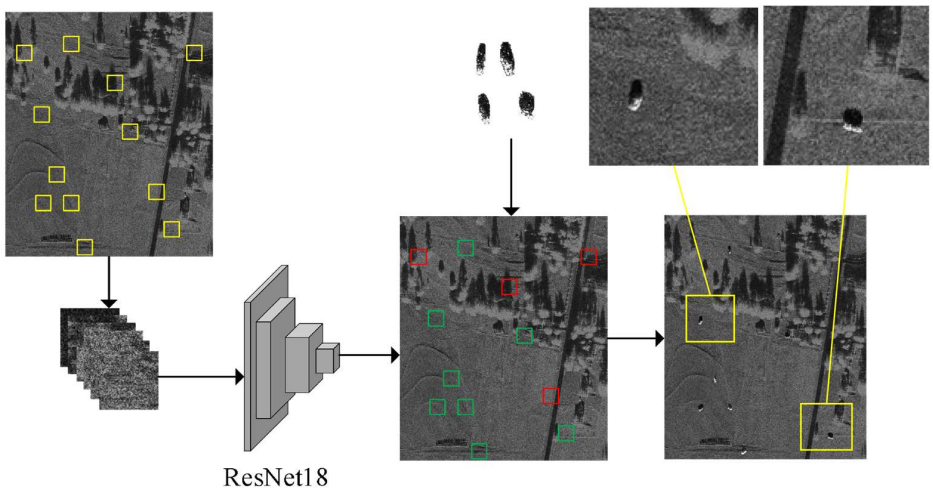


Figure 3. Schematic diagram of target subject placement methods. The yellow rectangles in the left upper image are the region input for background region classification. The green and red rectangles in the middle image are the placeable and non-placeable background regions, respectively. The right image is the final synthetic SAR image.

Afterwards, the Alpha blending technique is used to integrate the target object with the background image. Alpha blending can effectively control the transparency between the target object and the background, thereby achieving a smooth transition. Specifically, the process of the Alpha blending technique can be expressed as:

$$I_{fuse}(x,y) = \alpha I_{fg}(x,y) + (1-\alpha) I_{bg}(x,y), \quad (1)$$

where $I_{fg}(x,y)$, $I_{bg}(x,y)$ and $I_{fuse}(x,y)$ denote the pixel value of the target object, the background image and the fused image. α is the transparency (Alpha value) of the target image.

Through the background region classification and target–background fusion, the rationality of target placement is effectively improved, providing solid data support and methodological foundation for subsequent target detection.

3.2. Target Detection Model

To enhance the feature modeling capability and detection accuracy of the target detection network for complex SAR scenes while maintaining the network lightweight and detection speed, we improve YOLOv8-s [7] target detector. Specifically, a Spatial Feature Enhancement Module (SFEM) and a Gather and Distribute (GD) mechanism are developed.

3.2.1. Spatial Feature Enhancement Module

To achieve effective dynamic recalibration of features, inspired by classical attention mechanisms such as SE [31] and CBAM [33], we design the Spatial Feature Enhancement Module (SFEM), which is illustrated in Figure 4. The main idea of SFEM is to dynamically strengthen the discriminative power of feature maps by parallel modeling of local and global spatial contextual information. Structurally, SFEM adopts a channel-wise split-processing strategy combined with parallel convolution operations of multiple receptive fields, effectively improving the network ability to perceive spatial details and model contextual information. This design is particularly suitable for detecting small targets in SAR images.

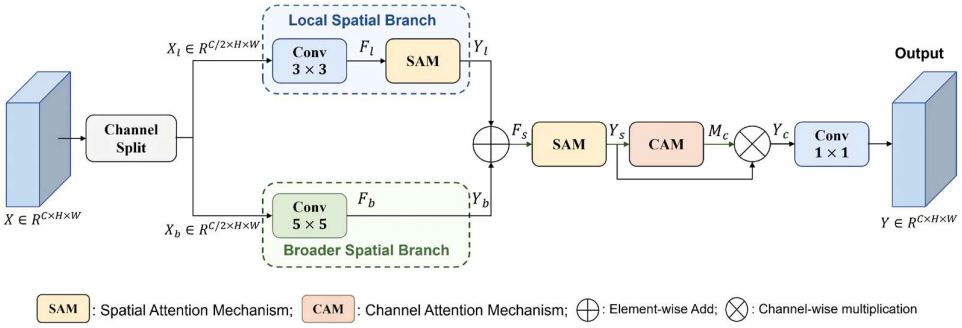


Figure 4. Architecture of the Spatial Feature Enhancement Module (SFEM). The input feature X is first split into two sub-features X_l and X_b, which are processed by local and broader context branches, respectively. The two branch outputs are fused by element-wise addition, refined by spatial attention and channel attention, and finally projected by a 1×1 convolution to generate the output feature Y with the same dimension as X.

Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, it is first split along the channel dimension into two sub-feature maps, $X_l, X_b \in \mathbb{R}^{C/2 \times H \times W}$. These two sub-features are then processed in parallel to capture multi-scale spatial contexts.

Local Context Branch (Top): X_l passes through a convolution layer (3×3) to extract intermediate local features F_l. A Spatial Attention Mechanism (SAM) is then applied to emphasize salient spatial details:

$$F_l = Conv_{3\times3}(X_l),$$

$$Y_l = F_l \odot \sigma(Conv([AvgPool(F_l); MaxPool(F_l)])), \tag{2}$$

where $Conv_{3\times3}(\cdot)$ represents a 3×3 convolution operation. $AvgPool(\cdot)$ and $MaxPool(\cdot)$ denote channel-wise average pooling and max pooling, respectively. $\sigma(\cdot)$ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication. Y_l is the local-context output after spatial attention refinement.

Broader Context Branch (Bottom): X_b passes through a parallel convolution layer (5×5) to capture broader contextual information:

$$Y_b = Conv_{5\times5}(X_b), \tag{3}$$

where $Conv_{5\times5}(\cdot)$ represents a 5×5 convolution operation, and Y_b is the output feature of the broader-context branch.

Feature Fusion and Refinement: The outputs from the two branches are fused via element-wise addition. To recalibrate the spatial importance across the aggregated multi-scale features, the summed feature map F_s undergoes a second Spatial Attention Mechanism (SAM), yielding the refined feature Y_s:

$$F_s = Y_l + Y_b,$$

$$Y_s = F_s \odot \sigma(Conv([AvgPool(F_s); MaxPool(F_s)])), \tag{4}$$

where the second SAM further estimates the spatial attention map from F_s, and Y_s represents the spatially refined multi-scale feature.

Channel Attention Mechanism (CAM): Finally, Y_s is fed into a Channel Attention Mechanism (CAM) to recalibrate the importance of different feature channels. The channel

attention weight M_c is generated by applying average pooling and max pooling followed by shared fully connected layers:

$$M_c = \sigma(FC(AvgPool(Y_s)) + FC(MaxPool(Y_s))), \tag{5}$$

where $FC(\cdot)$ represents the shared fully connected layers used to generate channel descriptor.

The attention weight M_c is then multiplied with Y_s along the channel dimension to obtain the channel-recalibrated feature:

$$Y_c = Y_s \otimes M_c, \tag{6}$$

where $\otimes$ denotes channel-wise multiplication.

Since the preceding channel split and element-wise fusion reduce the channel dimension to $C/2$, a 1×1 convolution ($Conv_{1\times1}$) is further applied to restore the channel dimension and produce the final output feature:

$$Y = Conv_{1\times1}(Y_c), \tag{7}$$

In SAR images, targets exhibit significant variations under different depression angles. In particular, at larger depression angles, target distortion and background noise make target detection more challenging. Through its local and global scale spatial attention mechanism, our SFEM effectively adapts to changes in target size and spatial relationships across different depression angles in SAR imagery. The subsequent channel attention fusion further enables precise recalibration of the importance of each feature channel, enhancing informative features while suppressing distracting ones.

3.2.2. Gather-And-Distribute Mechanism

For information fusion in the detection network, information loss may still occur during cross-layer feature interactions. To alleviate this problem, we introduce a Gather-and-Distribute (GD) mechanism, which aggregates multi-level features and redistributes the fused representation to different detection layers. The GD mechanism consists of two stages, the Gather stage and the Distribute stage, as illustrated in Figure 5 and Figure 6, respectively.

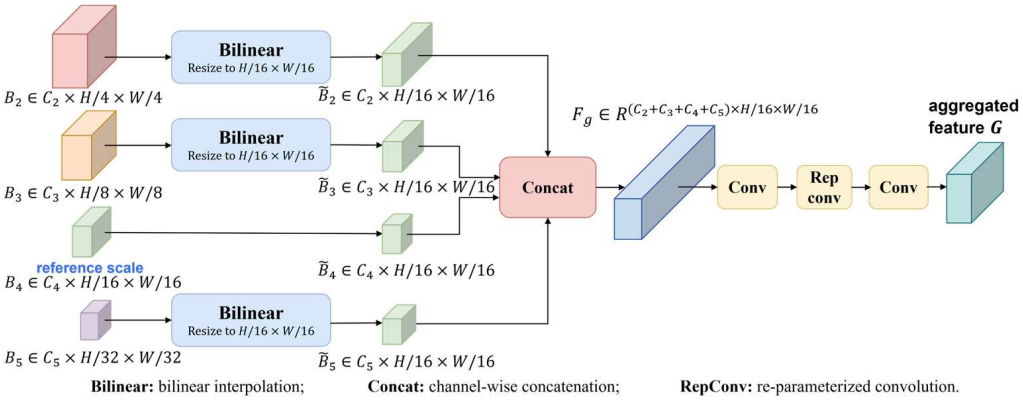


Figure 5. Architecture of the Gather module. The input features B_2, B_3, B_4, and B_5 are downsampled by factors of 4, 8, 16, and 32 with respect to the input image, respectively. B_4 is used as the reference scale, and the other features are resized to $H_4 \times W_4$ by bilinear interpolation. The aligned features $\widetilde{B}_2 - \widetilde{B}_5$ are concatenated along the channel dimension to form F_g, and a Conv–RepConv–Conv block further fuses F_g into the aggregated feature G.

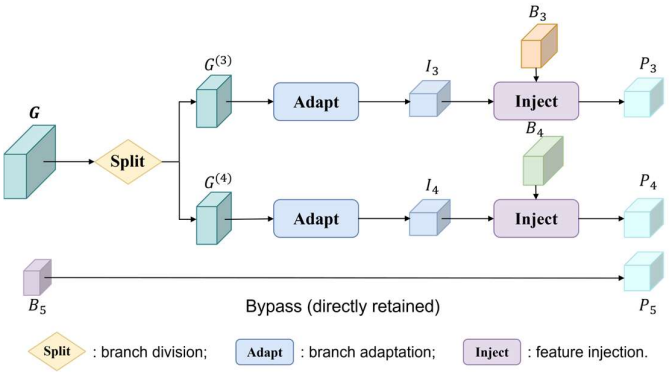


Figure 6. Architecture of the Distribute module. The aggregated feature G is split into two branches $G^{(3)}$ and $G^{(4)}$. Each branch is adapted to the spatial size and channel dimension of the corresponding feature level, producing I_3 and I_4. The adapted features are then injected into B_3 and B_4 through feature fusion operations to generate P_3 and P_4, respectively. The high-level feature B_5 is directly preserved through a bypass connection as P_5.

Let $\{B_i\}_{i=2}^{5}$ denote the input feature maps from different network stages, which are defined as

$$B_i \in \mathbb{R}^{C_i \times H_i \times W_i}, \tag{8}$$

As shown in Figure 5, B_2, B_3, B_4, and B_5 have spatial resolutions of 1/4, 1/8, 1/16, and 1/32 of the input image, respectively. To effectively fuse multi-scale features, B_4 is selected as the reference scale. The feature maps from the other stages are resized to the same spatial resolution as B_4 using bilinear interpolation:

$$\widetilde{B}_i = \mathcal{R}_{i \to 4}(B_i), i \in \{2,3,5\}, \tag{9}$$

where $\mathcal{R}_{i \to 4}(\cdot)$ denotes the bilinear interpolation operation that resizes the feature map from the i-th stage to the spatial resolution of B_4. After this operation, B_2, B_3, and B_5 are spatially aligned with B_4 for subsequent feature concatenation.

For the reference feature map, no resizing is applied:

$$\widetilde{B}_4 = B_4, \tag{10}$$

Then, the aligned multi-level features are concatenated along the channel dimension:

$$F_g = \text{Concat}\left(\widetilde{B}_2, \widetilde{B}_3, \widetilde{B}_4, \widetilde{B}_5\right), \tag{11}$$

where $\text{Concat}(\cdot)$ denotes channel-wise concatenation, and F_g represents the concatenated multi-level feature that integrates information from different network depths.

The concatenated feature F_g contains both low-level spatial details and high-level semantic information. However, direct concatenation may introduce channel redundancy and mixed scale-level responses. Therefore, a convolutional fusion block consisting of Conv–*RepConv*–Conv is adopted to compress redundant channels and enhance feature representation:

$$G = \phi_3\left(RepConv\left(\phi_2\left(F_g\right)\right)\right), \tag{12}$$

where $\phi_2(\cdot)$ and $\phi_3(\cdot)$ denote standard convolution operations, $RepConv(\cdot)$ denotes the re-parameterized convolution module, and G is the aggregated feature produced by the Gather stage.

Through this process, the Gather module integrates features from different network depths at a unified spatial scale. Low-level features provide fine-grained edge, texture,

and localization cues, while high-level features contribute stronger semantic representations. After Conv–*RepConv*–Conv fusion, the aggregated feature G can simultaneously preserve spatial details for target localization and semantic context for background suppression, thereby providing a more informative global representation for the subsequent Distribute stage.

As shown in Figure 6, after obtaining the aggregated feature G, the Distribute module first splits it into two branches:

$$\left[G^{(3)}, G^{(4)}\right] = \text{Split}(G), \tag{13}$$

where $\text{Split}(\cdot)$ denotes the channel-wise split operation.

The two branches G_3 and G_4 are then adapted to the corresponding feature levels, producing I_3 and I_4, which are injected into B_3 and B_4, respectively:

$$I_l = \mathcal{A}_l\left(G^{(l)}\right), l \in \{3,4\}, \tag{14}$$

$$P_l = \mathcal{J}_l(B_l, I_l), l \in \{3,4\}, \tag{15}$$

where $\mathcal{A}_l(\cdot)$ denotes the branch adaptation operation that adjusts G_l to the spatial size and channel dimension of the corresponding feature level. I_l is the adapted feature for the l-th detection layer. $\mathcal{J}_l(\cdot)$ denotes the Inject operation that fuses the original feature B_l with the adapted contextual feature I_l, and P_l is the enhanced output feature after injection.

The highest-level feature is directly retained as:

$$P_5 = B_5, \tag{16}$$

Therefore, the output multi-scale features of the Distribute stage are expressed as:

$$\{P_3, P_4, P_5\} = \{\mathcal{J}_3(B_3, I_3), \mathcal{J}_4(B_4, I_4), B_5\}, \tag{17}$$

where P_3, P_4, and P_5 denote the final output features of the Distribute stage at different scales. These features are used by the subsequent detection head for multi-scale SAR target detection.

Through this design, the Distribute module injects the global representation obtained from multi-level feature aggregation back into the key detection layers. Since B_3 retains more spatial details and localization cues, the enhanced feature P_3 is beneficial for small SAR target localization. Meanwhile, B_4 provides a better balance between spatial resolution and semantic abstraction, allowing P_4 to incorporate both contextual and discriminative information. By injecting the adapted contextual features into B_3 and B_4, the network can preserve local details while incorporating global semantic information, thereby improving the discriminability between target regions and background clutter.

Compared with the layer-by-layer propagation in conventional FPN/PAN structures, the Distribute module redistributes the global aggregated feature G, obtained from B_2, B_3, B_4, and B_5, to the key detection layers through split, adaptation, and injection operations. This design shortens the cross-scale information propagation path and reduces information attenuation caused by long cross-scale interactions. For SAR target detection, where targets are usually small and easily affected by speckle noise, shadow variations, and background clutter, the Distribute module enhances the responses of true target regions while suppressing irrelevant background activations, leading to improved robustness under complex scenes and varying depression angles.

The information gathering stage ensures that features of different scales are thoroughly fused, while the information distribution stage allocates these features appropriately so

that each layer gains access to global contextual information. This innovative feature fusion strategy enables YOLOv8 to more precisely capture information across scales from coarse to fine, thereby enhancing its ability to detect targets, particularly small ones. Finally, the overall network architecture is shown in Figure 7.

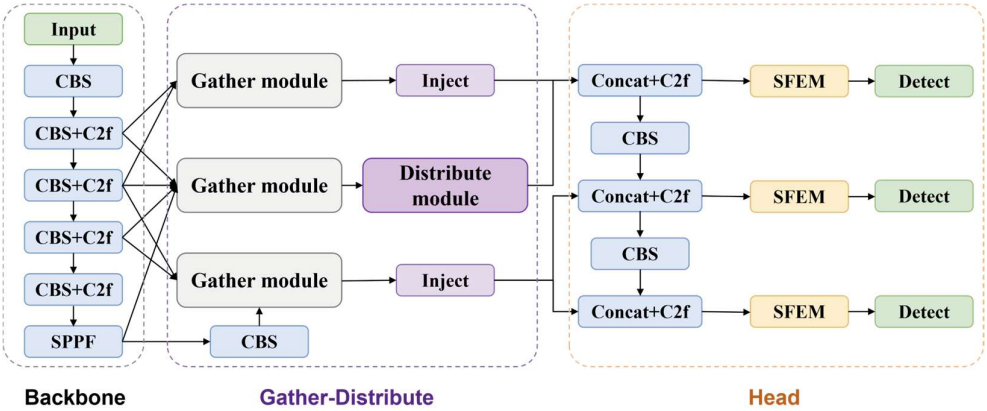


Figure 7. Overall structure of the proposed network.

3.2.3. Loss Function of DA-GDNet Detector

To train the proposed DA-GDNet detector, we adopt the same multi-task detection loss as the YOLOv8-based detector. It consists of three components: bounding-box localization loss, classification loss, and distribution focal loss. The overall detection loss is formulated as:

$$\mathcal{L}_{det} = \lambda_{box}\mathcal{L}_{box} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dfl}\mathcal{L}_{dfl}, \tag{18}$$

where $\mathcal{L}_{box}$, $\mathcal{L}_{cls}$, and $\mathcal{L}_{dfl}$ denote the bounding-box localization loss, classification loss, and distribution focal loss, respectively. The coefficients λ_{box}, λ_{cls}, and λ_{dfl} are used to balance the contributions of different loss terms.

For bounding-box localization, Complete IoU (CIoU) loss is adopted to measure the mismatch between the predicted bounding box and the ground-truth bounding box:

$$\mathcal{L}_{box} = \sum_{i=1}^{N_{pos}} \left(1 - \mathrm{CIoU}\left(b_i, b_i^{gt}\right)\right), \tag{19}$$

where N_{pos} is the number of positive samples, and b_i and b_i^{gt} denote the predicted bounding box and the corresponding ground-truth box, respectively.

The classification loss is implemented using BCEWithLogitsLoss, which combines sigmoid activation with binary cross-entropy. It is defined as follows:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N}\sum_{c=1}^{C}[y_{ic}\log(\sigma(s_{ic})) + (1 - y_{ic})\log(1 - \sigma(s_{ic}))], \tag{20}$$

where N is the number of prediction samples, C is the number of target categories, s_{ic} is the classification logit of the i-th prediction for class c, $\sigma(\cdot)$ denotes the sigmoid function, and y_{ic} is the corresponding ground-truth class label.

In addition, Distribution Focal Loss (DFL) is used to refine bounding-box regression by modeling the bounding-box offsets as discrete probability distributions. It can be expressed as:

$$\mathcal{L}_{dfl} = -\sum_{i=1}^{N_{pos}}[w_l \log P_i(l) + w_r \log P_i(r)], \tag{21}$$

where $P_i(l)$ and $P_i(r)$ denote the predicted probabilities of the two adjacent discrete bins around the target regression value, and w_l and w_r are the corresponding interpolation weights. By modeling the bounding-box offsets in a distributional form, DFL im-

proves localization precision, especially for small SAR targets with weak contours and cluttered surroundings.

3.3. Datasets

We conduct experiments on target and background images from the MSTAR dataset [13] under SOC, EOC-30 and EOC-45. The MSTAR dataset includes SAR images of various military ground vehicles and accounts for complex extended operating conditions such as target occlusion and configuration variations. In addition, it includes a series of scene images (Figure 8), each covering an area of approximately 0.1 km^2. For each depression angle, there are 50 images, consisting of 45 village scenes and 5 urban scenes. The village scenes primarily contain natural landscapes such as lakes, forests, grasslands, and farmland, whereas the urban scenes are dominated by man-made structures. The specific dataset configuration is summarized in Table 1.

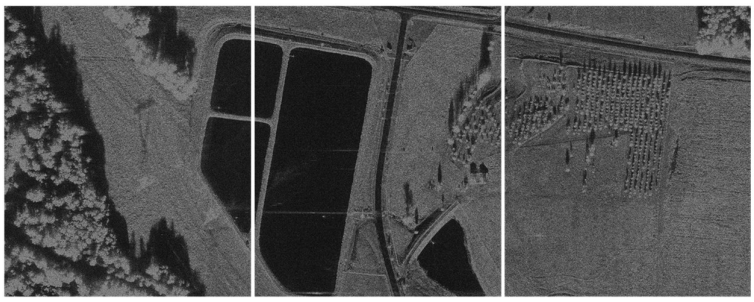


Figure 8. Scene images of MSTAR dataset.

Table 1. Data distribution of MSTAR dataset.

Class	15°	17°	30°	45°
2S1	274	299	288	303
BRDM_2	274	298	402	423
BTR_60	195	256	—	—
D7	274	299	—	—
SN_132	232	196	—	—
SN_9563	233	195	—	—
SN_C71	233	196	—	—
T62	273	299	—	—
ZIL131	274	299	—	—
ZSU_23_4	274	299	406	422

Target detection experiments are conducted under both Standard Operating Condition (SOC) and Extended Operating Condition (EOC). The composition of training and testing data for SOC and EOC is detailed in Tables 2 and 3.

Under the SOC experimental setting, the training and testing sets share highly consistent imaging conditions, such as identical depression angles and similar target configurations. This controlled environment allows us to isolate external interference and independently evaluate the effectiveness of the network architecture in feature extraction and classification.

The purpose of the EOC setting is to simulate the challenge when significant variations in radar depression angles occur. In this configuration, the training images are collected at a 17° depression angle, while the testing images are acquired at two larger depression angles, namely 30° and 45°, corresponding to EOC-30 and EOC-45.

Table 2. Detailed information of training and testing sets under SOC.

Class	Serial Number	Training	Testing
2S1	B01	299	274
BRDM_2	E71	298	274
BTR_60	7532	256	195
D7	13,015	299	274
T72	SN_132	232	196
BMP_2	SN_9563	233	195
BTR_70	SN_C71	233	196
T62	A51	299	273
ZIL131	E12	299	274
ZSU_23_4	D08	299	274

Table 3. Detailed information of training and testing sets under EOC-30 and EOC-45.

Class	Serial Number	Training	EOC-30°	EOC-45°
2S1	B01	299	288	303
BRDM_2	E71	298	287	303
ZSU_23_4	D08	299	288	303

To further evaluate the robustness and generalization ability of the proposed detector, we introduce an 8-class subset of the ATRNet-STAR dataset as an additional benchmark [15]. Compared with the MSTAR dataset, ATRNet-STAR contains a broader range of civilian and engineering vehicle targets with more diverse structural characteristics. Therefore, it provides a useful supplement for evaluating whether the proposed network can maintain stable detection performance across more target categories and under different imaging conditions. The selected classes include sedans, SUVs, vans, trucks, buses, and engineering vehicles, which increases the diversity of target size, shape, and scattering characteristics. The sample distribution of the selected ATRNet-STAR subset under different depression angles is summarized in Table 4.

Table 4. Data distribution of the 8-class ATRNet-STAR dataset.

Class	15°	30°	45°	60°
Buick_Excelle_GT	401	536	576	576
Chery_qq3	528	624	661	648
Jeep_Patriot	384	576	568	576
Buick_GL8	596	618	640	660
Dongfeng_Duolika	480	520	480	480
FAW_J6P	540	576	543	576
Yutong_ZK6120HY1	660	572	576	576
Hyundai_HLF25_II	536	556	564	576

In this study, the selected 8-class ATRNet-STAR subset was evaluated under the standard operating condition (SOC) and the extended operating condition at 30° (EOC-30). In the SOC setting, both the training and testing sets contain samples acquired at multiple depression angles, including 15°, 30°, 45°, and 60°. Therefore, this setting was used to evaluate the in-distribution detection performance of the model when the training and testing data share the same angle range and target categories. The per-class sample distribution of the SOC training and testing sets is reported in Table 5.

Table 5. SOC setting of the 8-class ATRNet-STAR dataset.

Class	Training	Testing
Buick_Excelle_GT	1436	653
Chery_qq3	1689	772
Jeep_Patriot	1441	663
Buick_GL8	1724	790
Dongfeng_Duolika	1392	568
FAW_J6P	1550	685
Yutong_ZK6120HY1	1646	738
Hyundai_HLF25_II	1577	655

The EOC-30 setting was used to evaluate the robustness of the detector under depression-angle variation. In this setting, the model was trained using samples acquired at a 15° depression angle and tested on samples acquired at a 30° depression angle. Compared with SOC, EOC-30 introduces an explicit angular domain shift between the training and testing data, making it suitable for assessing the generalization ability of the detector under changed imaging geometry. The per-class sample distribution of the EOC-30 training and testing sets is shown in Table 6.

Table 6. EOC setting of the 8-class ATRNet-STAR dataset.

Class	Training	Testing
Buick_Excelle_GT	401	536
Chery_qq3	528	624
Jeep_Patriot	384	576
Buick_GL8	596	618
Dongfeng_Duolika	480	520
FAW_J6P	540	576
Yutong_ZK6120HY1	660	572
Hyundai_HLF25_II	536	556

3.4. Experimental Settings

All experiments are implemented with the PyTorch 2.1.2 deep learning framework and conducted on RTX 2080Ti GPU with 12 GB memory. The input image size is set to 1024×1024, the batch size is 16, and the initial learning rate is 0.01, with cosine annealing learning rate scheduler. We adopt SGD as the optimizer with a momentum of 0.937 and weight decay of 5×10^{-4}, and the total number of training epochs is set to 200.

3.5. Evaluation Metrics

To quantitatively evaluate the detection performance of different methods, precision, recall, F1-score, average precision (AP), and mean average precision (mAP) are used as the main evaluation metrics. A predicted bounding box is considered a true positive (TP) when its intersection over union (IoU) with the corresponding ground-truth box is greater than a predefined threshold. Otherwise, it is regarded as a false positive (FP). A ground-truth target that is not correctly detected is counted as a false negative (FN). The IoU between the predicted box B_p and the ground-truth box B_g is defined as

$$IoU = \frac{B_p \cap B_g}{B_p \cup B_g}. \tag{22}$$

Precision measures the proportion of correctly detected targets among all predicted targets, while recall measures the proportion of correctly detected targets among all ground-truth targets. They are calculated as

$$Precision = \frac{TP}{TP + FP}. \tag{23}$$

The $F1$-score provides a balanced measurement of precision and recall, and is defined as

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{24}$$

The precision-recall curve was obtained by varying the confidence threshold. The average precision (AP) of each category was calculated as the area under the precision-recall curve:

$$AP = \int_0^1 P(R)dR, \tag{25}$$

where P and R denote precision and recall, respectively. For multi-class target detection, mAP was calculated by averaging AP values over all target categories:

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i, \tag{26}$$

where C is the number of target categories and AP_i represents the AP of the i-th category. In this study, mAP@0.5 is used to evaluate detection accuracy when the IoU threshold was set to 0.5. In addition, mAP@0.5 : 0.95 is reported when necessary to provide a stricter evaluation under multiple IoU thresholds from 0.5 to 0.95 with a step size of 0.05.

To further evaluate model efficiency, the number of parameters, giga floating-point operations (GFLOPs), and inference speed are also reported. The number of parameters reflects the model size, GFLOPs measure the computational complexity, and frames per second (FPS) indicates the inference efficiency. These metrics allow a more comprehensive comparison between detection accuracy and computational cost, which is important for practical SAR target detection applications.

4. Results

This section presents the experimental results of the proposed DA-GDNet. First, ablation studies are conducted to analyze the individual contributions of the data augmentation strategy, SFEM, and GD mechanism. Then, the overall detection performance on the MSTAR and ATRNet-STAR datasets is compared with several representative detection methods. Finally, qualitative visualization and feature visualization results are provided to further evaluate the effectiveness and interpretability of the proposed method.

4.1. Ablation Study

In this section, we conduct two groups of ablation studies. First, we analyze the effectiveness of different augmentation components to verify the contribution of the proposed augmentation strategy. Then, we evaluate the individual contributions of Data Augmentation (DA), the Spatial Feature Enhancement Module (SFEM), and the Gather-and-Distribute (GD) mechanism to the overall DA-GDNet framework.

4.1.1. Ablation Study on Augmentation Components

To analyze the contribution of each augmentation component, we conduct an ablation study on the MSTAR dataset under the SOC-15 setting. The compared settings include no augmentation, standard image-level augmentation, random copy-paste, SARBake-guided

U-Net extraction, SARBake + U-Net with background classification only, SARBake + U-Net with fusion only, and the complete proposed DA strategy. This experiment is designed to verify whether the performance improvement came from physically guided target/shadow extraction, target–background suitability assessment, and target–background fusion, rather than simply increasing the number of training samples. The results are shown in Table 7.

Table 7. Ablation study on augmentation components on the MSTAR dataset. ↑ indicates that higher values denote better performance.

Method	Precision ↑	Recall ↑	mAP@50 ↑	mAP@50:95 ↑
Baseline	71.4	79.5	86.6	72.9
Standard Augmention	78.6	83.2	90.4	77.5
Random Copy-Paste	83.4	86.5	92.6	80.9
SARBake + U-Net	88.2	89.7	95.6	84.9
SARBake + U-Net + background classification only	89.1	91.7	96.2	86.6
SARBake + U-Net + fusion only	90.6	90.3	97.5	85.6
Proposed DA	**91.5**	**91.8**	**97.8**	**87.3**

As shown in Table 7, all augmentation strategies improve the detection performance compared with the baseline. Standard augmentation increases mAP@50 from 86.6% to 90.4%, indicating that basic image-level transformations can enhance appearance diversity. Random copy-paste further improves mAP@50 to 92.6% by enriching target–background combinations. SARBake + U-Net achieves 95.6% mAP@50 and 84.9% mAP@50:95, showing that target/shadow-guided extraction is more effective than simple rectangular copy-paste. After introducing background classification alone, the mAP@50 and mAP@50:95 further increase to 96.2% and 86.6%, respectively, indicating that background suitability screening help generate more reasonable composite samples. The fusion-only variant achieves 97.5% mAP@50, suggesting that target–background fusion enhance target visibility and detection confidence. The complete proposed DA strategy achieves the best overall performance, with 91.5% precision, 91.8% recall, 97.8% mAP@50, and 87.3% mAP@50:95. These results demonstrate that the proposed DA strategy benefits from the joint use of physically guided target/shadow extraction, background-aware sample selection, and realistic fusion, rather than from data quantity expansion alone.

4.1.2. Ablation Study on Network Components

To further evaluate the effectiveness of the proposed DA-GDNet framework, we conduct a component-wise ablation study by gradually adding DA, SFEM, and the GD mechanism to the baseline detector. Unlike the previous augmentation component analysis, this experiment verifies the contribution of each major module to the complete detection framework. Experiments on both MSTAR and ATRNet-STAR under SOC and EOC settings are used to examine whether the proposed data-level and architecture-level modules consistently improve detection accuracy and robustness. The results are reported in Table 8 and Table 9, respectively.

For the MSTAR dataset, DA is the main source of improvement under SOC, increasing mAP@50 and mAP@50:95 from 86.6% and 72.9% to 97.8% and 87.3%, respectively. When SFEM or GD is further combined with DA, the localization accuracy is improved, and the complete DA-GDNet achieves the best mAP@50 of 98.9% and mAP@50:95 of 90.3%. Although the precision of the complete model is slightly lower than that of DA + SFEM, it obtains the highest recall of 96.2%, indicating stronger target coverage. Under the more challenging EOC-30 and EOC-45 settings, DA alone provides limited or unstable gains, whereas combining DA with SFEM and GD produces more consistent improvements. In particular,

the complete model increases mAP@50 from 81.3% to 90.1% under EOC-30 and from 47.4% to 62.1% under EOC-45, demonstrating better robustness to depression-angle variations.

Table 8. Ablation experiments under different conditions on MSTAR dataset. ↑ indicates that higher values denote better performance.

	Method				Precision ↑	Recall ↑	mAP@50 ↑	mAP@50:95 ↑
	Base	DA	SFEM	GD				
SOC	√				71.4	79.5	86.6	72.9
	√	√			91.5	91.8	97.8	87.3
	√		√		78.5	85.0	85.8	78.8
	√			√	77.8	86.2	86.5	79.6
	√	√	√		**95.4**	95.3	98.8	**90.3**
	√	√		√	93.2	96.1	98.6	**90.0**
	√	√	√	√	93.4	**96.2**	**98.9**	**90.3**
EOC-30	√				69.2	83.7	81.3	53
	√	√			79.3	82.7	84.9	59.3
	√		√		72.8	84.2	81.0	54.1
	√			√	73.5	83.8	82.5	55.0
	√	√	√		75.9	85.6	86.2	65.6
	√	√		√	77.4	85.2	85.5	66.5
	√	√	√	√	**81.7**	**85.8**	**90.1**	**67.4**
EOC-45	√				48.1	67.5	47.4	19.7
	√	√			51.8	59	44.9	19.3
	√		√		49.2	68.1	47.5	19.5
	√			√	50.8	68.8	46.5	20.0
	√	√	√		53.7	69.7	56.8	38.2
	√	√		√	52.0	70.6	56.0	37.6
	√	√	√	√	**58.1**	**71.4**	**62.1**	**38.6**

Table 9. Ablation experiments under different conditions on ATRNet-STAR dataset. ↑ indicates that higher values denote better performance.

	Method				Precision ↑	Recall ↑	mAP@50 ↑	mAP@50:95 ↑
	Base	DA	SFEM	GD				
SOC	√				95.8	94.8	97.3	62.7
	√	√			95.9	95.0	97.5	62.8
	√		√		95.0	94.9	97.3	62.6
	√			√	95.7	94.8	97.4	62.7
	√	√	√		95.8	**95.2**	97.5	62.9
	√	√		√	**96.1**	94.9	**97.6**	62.8
	√	√	√	√	96.0	95.0	**97.6**	**63.0**
EOC-30	√				75.4	71.0	77.3	42.6
	√	√			77.2	72.8	79.3	43.6
	√		√		76.8	72.0	78.8	43.2
	√			√	76.5	72.2	78.5	43.0
	√	√	√		78.8	74.2	80.8	43.7
	√	√		√	78.7	73.9	80.5	43.8
	√	√	√	√	**79.6**	**74.5**	**81.1**	**44.1**

The ablation results on the ATRNet-STAR dataset further verify the generalization ability of the proposed framework in large-scale cluttered scenes. Under SOC, the baseline already achieves high performance, and the complete DA-GDNet slightly improves mAP@50 from 97.3% to 97.6% and mAP@50:95 from 62.7% to 63.0%. In contrast, the advantages became more evident under EOC-30, where the baseline mAP@50 and mAP@50:95 are 77.3% and 42.6%, respectively. DA, SFEM, and GD each improves the results to different

extents, while their combination achieves the best performance, with 79.6% precision, 74.5% recall, 81.1% mAP@50, and 44.1% mAP@50:95. These results are consistent with those on MSTAR, indicating that the proposed DA strategy, SFEM, and GD mechanism jointly improve detection robustness under complex backgrounds and angular domain shifts.

4.2. Comparison Experiments

In this section, we compare the proposed DA-GDNet with representative detection networks to evaluate its effectiveness for SAR target detection under different operating conditions. To provide a more comprehensive comparison, the selected baselines include mainstream general detectors, such as YOLOv8-s [7], YOLOv11-s [53], and RT-DETR-l [54], as well as SAR-oriented detection models, including DenoDet [55], SFS-Conv [56], Deformer-FPN [57], and DiffDet4SAR [58]. Experiments are conducted on the MSTAR dataset under SOC, EOC-30, and EOC-45 settings, and on the ATRNet-STAR dataset under SOC and EOC-30 settings. In addition to precision, recall, mAP@50, and mAP@50:95, the number of parameters, GFLOPs, and inference speed are also reported to quantify model complexity and real-time detection capability. The comparison results are reported in Table 10 and Table 11, respectively.

Table 10. Comparison of different target detection networks on the MSTAR dataset. ↑ indicates that higher values denote better performance.

Dataset	Method	Precision ↑	Recall ↑	mAP@50 ↑	mAP@50:95 ↑	Params(M)	GFLOPs	FPS
SOC	YOLOv8-s	71.4	79.5	86.6	72.9	11.1	28.5	1666.7
	DenoDet	90.2	91.3	95.4	83.1	34.2	12.9	215.0
	DiffDet4SAR	82.3	86.1	89.4	70.5	48.0	36.2	43.5
	YOLOv11-s	87.4	84.9	94.2	82.7	9.4	21.3	1666.7
	RT-DETR-l	**95.1**	95.1	95.9	84.8	32.0	103.5	588.2
	SFS-Conv	42.2	61.7	51.4	40.8	9.7	23.6	1250.0
	Deformer-FPN	94.9	94	98.3	84.5	9.1	31.0	172.6
	DA-GDNet(ours)	93.4	**96.2**	**98.9**	**90.3**	12.1	32.1	1250.0
EOC-30	YOLOv8-s	69.2	83.7	81.3	53.0	11.2	28.5	1250.0
	DenoDet	78.6	82.1	86.2	62.3	34.2	12.9	215.0
	DiffDet4SAR	72.5	78.4	81.4	54.7	48.0	36.2	43.5
	YOLOv11-s	78.1	85.4	84.2	59.4	9.4	21.3	1666.7
	RT-DETR-l	79.4	74	74.3	51.2	32.0	103.5	588.2
	SFS-Conv	74.1	77.3	84.8	62.7	9.7	23.6	1250.0
	Deformer-FPN	80.7	80.4	83.4	58.6	9.1	31.0	176.0
	DA-GDNet(ours)	**81.7**	**85.8**	**90.1**	**67.4**	12.1	32.1	1250.0
EOC-45	YOLOv8-s	48.1	67.5	47.4	19.7	11.1	28.5	1666.7
	DenoDet	53.4	67.0	54.6	28.9	34.2	12.9	215.0
	DiffDet4SAR	49.7	64.9	49.8	23.0	48.0	36.2	43.5
	YOLOv11-s	57.2	66.9	51.0	23.0	9.4	21.3	1666.7
	RT-DETR-l	51.9	46.2	37.1	19.9	32.0	103.5	588.2
	SFS-Conv	37.2	**71.8**	39.7	17.4	9.7	23.6	1250.0
	Deformer-FPN	51.7	65.7	52	26.6	9.1	31.0	172.6
	DA-GDNet(ours)	**58.1**	71.4	**62.1**	**38.6**	12.1	32.1	1250.0

For the MSTAR dataset, DA-GDNet achieves the best overall detection performance under all operating conditions. Under SOC, it obtains the highest mAP@50 and mAP@50:95, reaching 98.9% and 90.3%, respectively. Compared with the lightweight YOLOv8-s baseline, DA-GDNet increases the parameter size only from 11.1 M to 12.1 M and the computational cost from 28.5 to 32.1 GFLOPs, while improving mAP@50 by 12.3 percentage points and mAP@50:95 by 17.4 percentage points. When the depression angle changes to 30° and 45°, most detectors show clear performance degradation. By contrast, DA-GDNet consistently achieves the highest mAP@50 and mAP@50:95 under both EOC settings. It also maintains

1250 FPS, which is much faster than DenoDet, DiffDet4SAR, RT-DETR-l, and Deformer-FPN, demonstrating a favorable balance between accuracy, complexity, and inference efficiency.

Table 11. Comparison of different target detection networks on the ATRNet-STAR dataset. ↑ indicates that higher values denote better performance.

Dataset	Method	Precision ↑	Recall ↑	mAP@50 ↑	mAP@50:95 ↑	Params(M)	GFLOPs	FPS
SOC	YOLOv8-s	95.8	94.8	97.3	62.7	11.1	28.5	1666.7
	DenoDet	95.1	94.0	96.9	60.8	34.2	12.9	215.0
	DiffDet4SAR	89.5	88.8	90.6	52.0	48.0	36.2	43.5
	YOLOv11-s	94.2	93.2	96.6	60.5	9.4	21.3	1666.7
	RT-DETR-l	**96.4**	**95.3**	96.8	61.5	32.0	103.5	588.2
	SFS-Conv	80.5	77.1	83.3	46.6	9.7	23.6	1250.0
	Deformer-FPN	96.0	94.7	96.7	56.3	9.1	31.0	172.6
	DA-GDNet(ours)	96.0	95	**97.6**	**63**	12.1	32.1	1250.0
EOC-30	YOLOv8-s	75.4	71	77.3	42.6	11.2	28.5	1250.0
	DenoDet	77.4	73.5	79.0	43.5	34.2	12.9	215.0
	DiffDet4SAR	69.3	64.5	71.4	36.1	48.0	36.2	43.5
	YOLOv11-s	75.7	72.7	78.2	43.0	9.4	21.3	1666.7
	RT-DETR-l	76.2	74.8	78.0	**44.3**	32.0	103.5	588.2
	SFS-Conv	41.4	46.8	40.4	19.7	9.7	23.6	1250.0
	Deformer-FPN	**80.1**	72.8	78.7	41.9	9.1	31.0	176.0
	DA-GDNet(ours)	79.6	**74.5**	**81.1**	44.1	12.1	32.1	1250.0

For the ATRNet-STAR dataset, DA-GDNet also shows stable performance in large-scale cluttered scenes. Under SOC, although RT-DETR-l achieves slightly higher precision and recall, DA-GDNet obtains the best mAP@50 and mAP@50:95, with values of 97.6% and 63.0%, respectively. Compared with YOLOv8-s, DA-GDNet introduces only limited additional computation, while further improving localization accuracy. Under the more challenging EOC-30 setting, several baselines degrade noticeably, especially SFS-Conv and DiffDet4SAR. DA-GDNet achieves the highest mAP@50 of 81.1% and a competitive mAP@50:95 of 44.1%, only 0.2 percentage points lower than RT-DETR-l. Meanwhile, it uses 12.1 M parameters, 32.1 GFLOPs, and reaches 1250 FPS. These results indicate that the proposed SFEM and GD mechanism improve SAR target representation and robustness without introducing excessive model complexity.

Overall, DA-GDNet achieves a better balance among detection accuracy, model complexity, and inference speed than both mainstream general detectors and SAR-specific models. The results on MSTAR and ATRNet-STAR demonstrate that the proposed method maintains strong robustness and generalization under both SOC and EOC settings, while preserving lightweight and real-time detection capability.

4.3. Qualitative and Feature Visualization Analysis

To further analyze the effectiveness of SFEM, Grad-CAM-based [59] feature activation maps are compared between the networks with and without SFEM on representative cluttered SAR samples from the MSTAR-derived test set. Four images containing typical background interference, including vegetation textures, shadowed regions, and strong clutter around target scattering areas, were selected for visualization. The visualization results are shown in Figure 9.

As shown in Figure 9, the network without SFEM tends to produce scattered responses over background textures, especially in vegetation and shadowed clutter regions. After introducing SFEM, the activation responses become more concentrated around target-related scattering regions, and the responses to irrelevant clutter are visibly reduced. This visualization provides intuitive evidence that SFEM helps suppress background interference and enhances discriminative target features in complex SAR scenes.

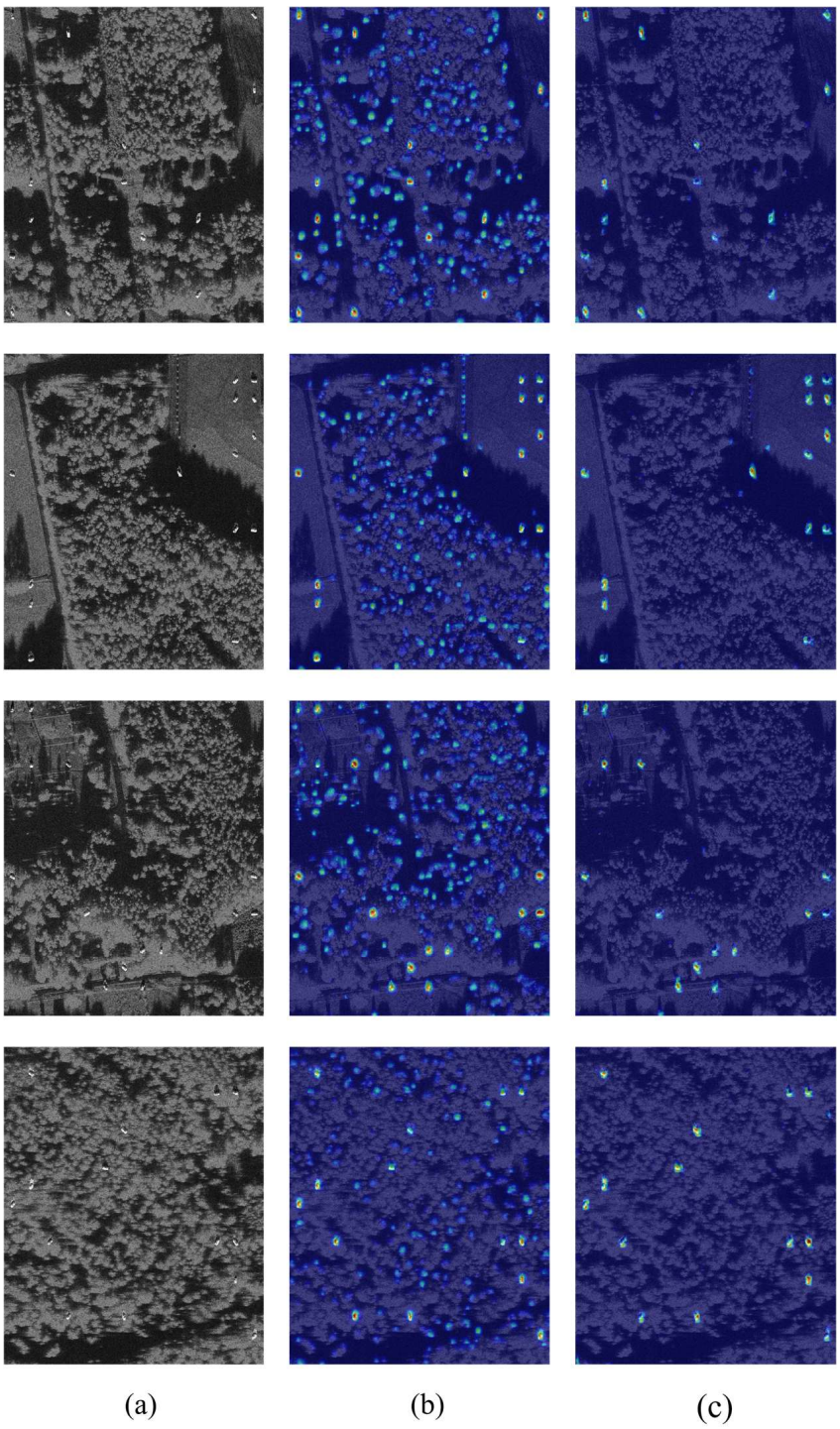


Figure 9. Grad-CAM visualization comparison between the baseline model without SFEM and with SFEM under strong-clutter SAR scenes. Columns (**a**–**c**) represent the original SAR images, Grad-CAM results of the baseline model without SFEM, and Grad-CAM results of the model with SFEM, respectively. Warmer colors indicate stronger feature responses, whereas cooler colors indicate weaker responses.

In addition, multi-scale feature activation maps are visualized to further examine the role of the GD mechanism in feature transmission. The activation maps are generated from the P3, P4, and P5 layers by averaging the absolute feature responses across channels and normalizing them for visualization. The visualization results are shown in Figure 10.

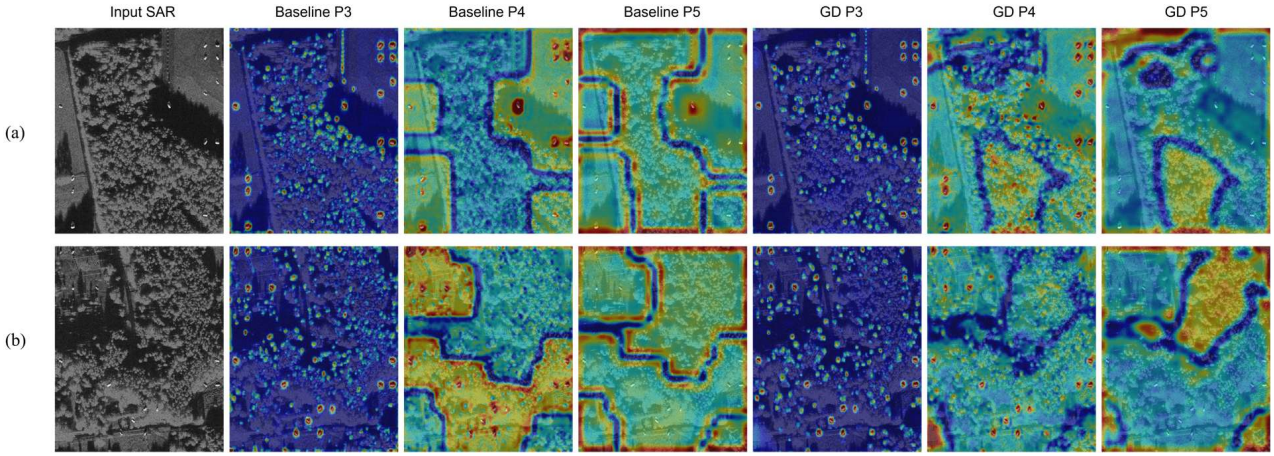


Figure 10. Multi-scale feature activation visualization of the baseline neck and the GD-based neck under cluttered SAR scenes. Subpanel (**a**) presents the first representative SAR test sample, whereas subpanel (**b**) presents the second representative SAR test sample. In each subpanel, the first column shows the input SAR image, the next three columns show the baseline feature activations at P3, P4, and P5, and the final three columns show the corresponding GD-based feature activations at P3, P4, and P5. The activation maps are obtained by averaging the absolute feature responses across channels and normalizing them for visualization. Warmer colors indicate stronger feature responses, whereas cooler colors indicate weaker feature responses.

As shown in Figure 10, the baseline neck produces relatively fragmented responses at the P3 level, while the responses at deeper P4 and P5 levels become more spatially coarse and are increasingly dominated by large background structures. In contrast, the GD-based neck maintains stronger target-related responses across multiple scales, especially at the P3 and P4 levels. These results indicate that the GD mechanism can alleviate information attenuation during cross-layer feature propagation and enhance multi-scale feature representation for small SAR targets in cluttered backgrounds.

5. Discussion

5.1. Effectiveness of Proposed Data Augmentation Strategy

At the data level, DA-GDNet introduces a structured data augmentation strategy to improve the realism and detection relevance of SAR training samples. Unlike standard image-level augmentation or random copy-paste, the proposed strategy follows a label-extract-screen-compose pipeline. SARBake-derived target and shadow priors provide geometric guidance for target extraction, U-Net segmentation refines real SAR target signatures, background classification avoids unreasonable target–background combinations, and target–background fusion reduces artificial boundaries in synthesized images. This design is intended to generate augmented samples that are not only more diverse, but also more consistent with SAR imaging characteristics.

The ablation results in Table 7 support this interpretation. Standard augmentation and random copy-paste improve detection performance to some extent, indicating that increasing sample diversity is useful. However, the larger improvement obtained by SARBake-guided extraction, background classification, fusion, and the complete DA strategy shows that sample quality is equally important. In particular, the progressive performance gains suggest that physically guided target extraction, contextual background screening, and smoother target–background integration jointly contribute to more effective SAR data augmentation.

5.2. Effectiveness of SFEM and GD

At the network level, SFEM and GD are designed to address two key difficulties in SAR target detection: weak small-target responses and information loss during multi-scale feature fusion. SFEM enhances target-related spatial responses by combining local spatial details with broader contextual information and then recalibrates features through spatial and channel attention. GD further gathers features from different network stages and redistributes the fused information to key detection layers, which helps preserve both semantic information and fine-grained localization cues.

The ablation results in Tables 8 and 9 show that SFEM and GD consistently improve detection performance under different experimental settings. These numerical improvements are also supported by the visualization results. Figure 9 shows that SFEM makes activation responses more concentrated around target regions and reduces interference from cluttered backgrounds. Figure 10 further indicates that GD produces clearer and more continuous target-related feature activations across detection layers. Therefore, the two modules improve the detector not only by increasing evaluation metrics, but also by producing more discriminative and stable feature representations.

5.3. Robustness Under Depression-Angle Variations and Complex Clutter

As a complete framework, DA-GDNet improves SAR target detection through the joint effect of data-level augmentation and feature-level enhancement. The proposed DA strategy expands plausible target–background distributions, while SFEM and GD strengthen target saliency, multi-scale feature interaction, and localization stability. This combination is particularly useful under depression-angle variations, because changes in imaging angle can alter target scattering patterns, shadow shapes, and apparent geometric structures, making SAR targets more difficult to distinguish from complex clutter.

The comparison results in Tables 10 and 11 demonstrate the robustness of DA-GDNet on both MSTAR and ATRNet-STAR. Under SOC settings, the proposed method maintains strong detection accuracy, showing its effectiveness under relatively standard test conditions. Under EOC settings, it also shows better adaptability to depression-angle changes and complex background interference than most compared detectors. These results indicate that DA-GDNet does not rely on a single dataset or a single experimental condition.

5.4. Ethical and Dual-Use Considerations

The present study is limited to academic research on synthetic aperture radar image analysis, data augmentation, and computer vision. MSTAR was used as an existing benchmark dataset to evaluate the proposed method under controlled experimental conditions. No new military data were collected, and the study did not involve real-time surveillance, geographic target localization, operational decision-making, weapons development or control, or deployment in an operational system. The additional experiments on the civilian and engineering vehicle classes of ATRNet-STAR were included to evaluate the method beyond the military-vehicle benchmark and to demonstrate its relevance to general vehicle detection in remote-sensing imagery.

The authors nevertheless acknowledge that research involving the detection and augmentation of military-vehicle imagery may have potential dual-use implications. Potential misuse could include adapting the method to support unauthorized surveillance or military target analysis. To mitigate these risks, the manuscript transparently identifies the military-vehicle content of MSTAR, restricts its claims to offline academic benchmark evaluation, does not claim readiness for operational deployment, and avoids providing operational targeting, geographic, or weapons-related information. The broader potential benefits of the underlying methodology include reducing annotation requirements and improving

image-analysis robustness in civilian applications such as transportation monitoring, infrastructure inspection, disaster assessment, and environmental observation. Any future use in sensitive or operational settings should be subject to separate institutional review, legal assessment, human oversight, and applicable national and international regulations.

6. Conclusions

In this paper, we propose DA-GDNet, a data-augmented SAR image target detection network, to address the limitations of unnatural target background integration and low adaptability to variations in depression angles. By combining SARBake-derived target/shadow priors with U-Net-based segmentation, the proposed method achieves fine-grained extraction of target and shadow regions. Then, background region classification and target–background fusion are designed to improve the target–background compatibility and contextual plausibility of the synthesized SAR images, thereby effectively enhancing data diversity and authenticity. For the detection network, the Gather-and-Distribute (GD) mechanism is introduced to achieve efficient multi-scale feature fusion, substantially enhancing the network's capability to detect multi-scale targets. In addition, the Spatial Feature Enhancement Module (SFEM) is designed to capture both local and global spatial contextual information, thereby improving the detection accuracy and stability under complex backgrounds and varying depression angles. The experimental results demonstrate that the proposed data augmentation method effectively increases the contextual plausibility and detection relevance of the augmented samples, while the improved detection model exhibits excellent generalization performance in complex scenarios and multi-view conditions.

Although DA-GDNet has achieved promising detection performance under different depression angles and complex backgrounds, several limitations still need to be further investigated. First, the proposed network introduces additional feature enhancement and multi-scale fusion modules, which may increase computational cost compared with the original lightweight detector. Therefore, future work will focus on further optimizing the network structure through model pruning, knowledge distillation, and lightweight attention design to improve its computational efficiency in offline academic benchmark evaluations. Second, the current physics-guided data augmentation pipeline mainly focuses on limited vehicle categories and target–background compositions. In the future, we will further evaluate the generalizability of the proposed augmentation strategy using additional civilian and engineering vehicle datasets and broader imaging conditions. Third, SAR target detection under extremely low signal-to-noise ratios, severe speckle noise, and strong background clutter remains challenging. Future research will explore joint denoising–detection models and uncertainty-aware feature learning to improve the robustness of DA-GDNet under degraded SAR imaging conditions. In addition, its applicability to civilian remote-sensing tasks, such as transportation monitoring, infrastructure inspection, disaster assessment, and environmental observation, will be further investigated. These future studies will remain subject to applicable ethical, legal, and institutional requirements.

Author Contributions: Conceptualization, F.Z. and Y.L.; methodology, F.Z.; software, F.Z.; validation, F.Z. and W.W.; formal analysis, F.Z.; investigation, F.Z.; resources, Y.L. and H.C.; data curation, F.Z. and W.W.; writing—original draft preparation, F.Z.; writing—review and editing, F.Z. and Y.S.; visualization, F.Z.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded in part by the Hebei Natural Science Foundation (No. F2026105031).

Data Availability Statement: This study used the existing MSTAR and ATRNet-STAR benchmark datasets cited in the manuscript. Access to and reuse of these datasets are subject to the conditions specified by their respective providers. No new military imagery, operational geographic information, or weapons-related data were collected or released in this study. The source code developed in this study is available from the corresponding author upon reasonable request.

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

References

1. Zhang, Y.; Zhu, D.; Mao, X.; Yu, X.; Zhang, J.; Li, Y. Multirotors video synthetic aperture radar: System development and signal processing. *IEEE Aerosp. Electron. Syst. Mag.* **2020**, *35*, 32–43. [CrossRef]
2. Wang, C.; Luo, S.; Pei, J.; Huang, Y.; Zhang, Y.; Yang, J. Crucial feature capture and discrimination for limited training data SAR ATR. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 291–305. [CrossRef]
3. Wu, B.; Liu, C.; Chen, J. A Review of Spaceborne High-Resolution Spotlight/Sliding Spotlight Mode SAR Imaging. *Remote Sens.* **2025**, *17*, 38. [CrossRef]
4. Zhang, Q.; Xiao, F.; Ding, Z.; Ke, M.; Zeng, T. Sliding spotlight mode imaging with GF-3 spaceborne SAR sensor. *Sensors* **2018**, *18*, 43. [CrossRef] [PubMed]
5. Mittermayer, J.; Wollstadt, S.; Prats-Iraola, P.; Scheiber, R. The TerraSAR-X staring spotlight mode concept. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3695–3706. [CrossRef]
6. Schwerdt, M.; Tous-Ramon, N.; Döring, B.; Zink, M.; Bachmann, M.; Schrank, D.; Gonzalez, J.H. Radiometric Performance of the TerraSAR-X Mission over More Than Ten Years of Operation. *Remote Sens.* **2018**, *10*, 754. [CrossRef]
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
8. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4203–4212. [CrossRef]
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
10. Shi, Y.; Du, L.; Guo, Y. Unsupervised domain adaptation for SAR target detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6372–6385. [CrossRef]
11. Wang, C.; Pei, J.; Wang, Z.; Huang, Y.; Wu, J.; Yang, H.; Yang, J. When deep learning meets multi-task learning in SAR ATR: Simultaneous target recognition and segmentation. *Remote Sens.* **2020**, *12*, 3863. [CrossRef]
12. Li, W.; Yang, W.; Liu, L.; Zhang, W.; Liu, Y. Discovering and explaining the noncausality of deep learning in SAR ATR. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
13. Keydel, E.R.; Lee, S.W.; Moore, J.T. MSTAR extended operating conditions: A tutorial. In *Proceedings of the Aerospace/Defense Sensing and Controls, Orlando, FL, USA, June 1996*; Zelnio, E.G., Douglass, R.J., Eds.; SPIE: Bellingham, WA, USA; pp. 228–242. [CrossRef]
14. Huang, L.; Liu, B.; Li, B.; Guo, W.; Yu, W.; Zhang, Z.; Yu, W. OpenSARShip: A dataset dedicated to Sentinel-1 ship interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 195–208. [CrossRef]
15. Liu, Y.; Li, W.; Liu, L.; Zhou, J.; Peng, B.; Song, Y.; Xiong, X.; Yang, W.; Liu, T.; Liu, Z.; et al. ATRNet-STAR: A large dataset and benchmark toward remote sensing object recognition in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 6735–6753. [CrossRef] [PubMed]
16. Li, Y.; Li, X.; Li, W.; Hou, Q.; Liu, L.; Cheng, M.-M.; Yang, J. SARDet-100K: Towards open-source benchmark and toolkit for large-scale SAR object detection. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; Advances in Neural Information Processing Systems. 37, pp. 128430–128461. [CrossRef]
17. Belli, M.I.; Xu, X. Affine and magnitude transforms based data augmentation comparison for SAR automatic target recognition. In Proceedings of the IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018; pp. 1649–1653. [CrossRef]
18. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 364–368. [CrossRef]
19. Zeng, Z.; Tan, X.; Zhang, X.; Huang, Y.; Wan, J.; Chen, Z. ATGAN: A SAR target image generation method for automatic target recognition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6290–6307. [CrossRef]

20. Wang, C.; Pei, J.; Liu, X.; Huang, Y.; Mao, D.; Zhang, Y.; Yang, J. SAR target image generation method using azimuth-controllable generative adversarial network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9381–9397. [CrossRef]
21. Oh, J.; Kim, M. PeaceGAN: A GAN-based multi-task learning method for SAR target image generation with a pose estimator and an auxiliary classifier. *Remote Sens.* **2021**, *13*, 3939. [CrossRef]
22. Ju, M.; Niu, B.; Hu, Q. SARGAN: A novel SAR image generation method for SAR ship detection task. *IEEE Sens. J.* **2023**, *23*, 28500–28512. [CrossRef]
23. Dvornik, N.; Mairal, J.; Schmid, C. On the importance of visual context for data augmentation in scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2014–2028. [CrossRef] [PubMed]
24. Furukawa, H. Deep learning for target classification from SAR imagery: Data augmentation and translation invariance. *arXiv* **2017**, arXiv:1708.07920. [CrossRef]
25. Liao, L.; Du, L.; Guo, Y. Semi-Supervised SAR Target Detection Based on an Improved Faster R-CNN. *Remote Sens.* **2022**, *14*, 143. [CrossRef]
26. Shen, L.; Tian, Y.; Zhang, L.; Yin, G.; Shuai, T.; Liang, S.; Wu, Z. Threshold Filtering Semi-Supervised Learning Method for SAR Target Recognition. *Comput. Mater. Contin.* **2022**, *73*, 465–476. [CrossRef]
27. Liao, L. Semi-Supervised SAR Target Detection with Cross-Domain Transfer Learning Based on YOLOv5. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Athens, Greece, 7–12 July 2024; pp. 7401–7404. [CrossRef]
28. Luo, C.; Zhang, Y.; Guo, J.; Pan, X.; Chen, X.; Gong, H. SAR-CDSS: A Semi-Supervised Cross-Domain Object Detection from Optical to SAR Domain. *Remote Sens.* **2024**, *16*, 940. [CrossRef]
29. Zhou, Z.; Cui, Z.; Tang, K.; Tian, Y.; Pi, Y.; Cao, Z. Gaussian Meta-Feature Balanced Aggregation for Few-Shot Synthetic Aperture Radar Target Detection. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 89–106. [CrossRef]
30. Wang, Z.; Yang, B.; Yue, H.; Ma, Z. Fine-Grained Prototypes Distillation for Few-Shot Object Detection. *arXiv* **2024**, arXiv:2401.07629. [CrossRef]
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [CrossRef]
33. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
34. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717. [CrossRef]
35. Deng, Y.; Guan, D.; Chen, Y.; Yuan, W.; Ji, J.; Wei, M. SAR-ShipNet: SAR-Ship Detection Neural Network via Bidirectional Coordinate Attention and Multi-Resolution Feature Fusion. *arXiv* **2022**, arXiv:2203.15480. [CrossRef]
36. Ma, X.; Cheng, J.; Li, A.; Zhang, Y.; Lin, Z. AMANet: Advancing SAR Ship Detection with Adaptive Multi-Hierarchical Attention Network. *arXiv* **2024**, arXiv:2401.13214. [CrossRef]
37. Chen, H.; Chen, C.; Wang, F.; Shi, Y.; Zeng, W. RSNet: A Light Framework for the Detection of Multi-Scale Remote Sensing Targets. *arXiv* **2024**, arXiv:2410.23073. [CrossRef]
38. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
39. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. [CrossRef]
40. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
41. Ke, X.; Zhang, X.; Zhang, T.; Shi, J.; Wei, S. SAR Ship Detection Based on Swin Transformer and Feature Enhancement Feature Pyramid Network. *arXiv* **2022**, arXiv:2209.10421. [CrossRef]
42. Zhao, C.; Fu, X.; Dong, J.; Cao, S.; Zhang, C. Enhancing, Refining, and Fusing: Towards Robust Multi-Scale and Dense Ship Detection. *arXiv* **2025**, arXiv:2501.06053. [CrossRef]
43. Ding, B.; Wen, G.; Huang, X.; Ma, C.; Yang, X. Data augmentation by multilevel reconstruction using attributed scattering center for SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 979–983. [CrossRef]
44. Cui, Z.; Zhang, M.; Cao, Z.; Cao, C. Image data augmentation for SAR sensor via generative adversarial nets. *IEEE Access* **2019**, *7*, 42255–42268. [CrossRef]

45. Xie, D.; Ma, J.; Li, Y.; Liu, X. Data augmentation of SAR sensor image via information maximizing generative adversarial net. In Proceedings of the IEEE 4th International Conference on Electronic Information and Communication Technology (ICEICT), Xi'an, China, 18–20 August 2021; pp. 454–458. [CrossRef]
46. Ju, M.; Niu, B.; Zhang, J. SAR image generation method for oriented ship detection via generative adversarial networks. *Signal Image Video Process.* **2024**, *18*, 589–596. [CrossRef]
47. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.-Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2917–2927. [CrossRef]
48. Geng, Z.; Xu, Y.; Wang, B.-N.; Yu, X.; Zhu, D.-Y.; Zhang, G. Target recognition in SAR images by deep learning with training data augmentation. *Sensors* **2023**, *23*, 941. [CrossRef] [PubMed]
49. Zhao, Z.; Xue, X.; Mariam, I.; Zhou, X. Integrating target and shadow features for SAR target recognition. *Sensors* **2023**, *23*, 8031. [CrossRef] [PubMed]
50. Malmgren-Hansen, D.; Nobel-Jorgensen, M. Convolutional neural networks for SAR image segmentation. In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Abu Dhabi, United Arab Emirates, 7–10 December 2015; pp. 231–236. [CrossRef]
51. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
53. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
54. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974. [CrossRef]
55. Dai, Y.; Zou, M.; Li, Y.; Li, X.; Ni, K.; Yang, J. DenoDet: Attention as deformable multisubspace feature denoising for target detection in SAR images. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 4729–4743. [CrossRef]
56. Li, K.; Wang, D.; Hu, Z.; Zhu, W.; Li, S.; Wang, Q. Unleashing channel potential: Space-frequency selection convolution for SAR object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 17323–17332. [CrossRef]
57. Yang, F.; Ouyang, T.; Liu, C. Deformer-FPN: FPN network based on deformable convolution and lightweight upsampling operator for ship detection in SAR image. In Proceedings of the IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Zhuhai, China, 22–24 November 2024; pp. 1–5. [CrossRef]
58. Zhou, J.; Xiao, C.; Peng, B.; Liu, Z.; Liu, L.; Liu, Y.; Li, X. DiffDet4SAR: Diffusion-based Aircraft Target Detection Network for SAR Images. *arXiv* **2024**, arXiv:2404.03595. [CrossRef]
59. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]

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