remote sensing

MDPI

Article

Aircraft Detection from Low SCNR SAR Imagery Using Coherent Scattering Enhancement and Fused Attention Pyramid

Xinzheng Zhang [1,2,*], Dong Hu [1], Sheng Li [3], Yuqing Luo [1], Jinlin Li [1] and Ce Zhang [4,5]

[1] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China; 20163970@cqu.edu.cn (D.H.); 20173802@cqu.edu.cn (Y.L.); 202212021005@stu.cqu.edu.cn (J.L.)
[2] Chongqing Key Laboratory of Space Information Network and Intelligent Information Fusion, Chongqing 400044, China
[3] Science and Technology on Electromagnetic Scattering Laboratory, Beijing 100854, China; lisheng2008@sina.com
[4] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK; c.zhang9@lancaster.ac.uk
[5] UK Centre for Ecology & Hydrology, Library Avenue, Lancaster LA1 4AP, UK
* Correspondence: zhangxinzheng@cqu.edu.cn

Abstract: Recently, methods based on deep learning have been applied to target detection using synthetic aperture radar (SAR) images. However, due to the SAR imaging mechanism and low signal-clutter-noise-ratio (SCNR), it is still a challenging task to perform aircraft detection using SAR imagery. To address this issue, a novel aircraft detection method is proposed for low SCNR SAR images that is based on coherent scattering enhancement and a fusion attention mechanism. Considering the scattering characteristics discrepancy between human-made targets and natural background, a coherent scattering enhancement technique is introduced to heighten the aircraft scatter information and suppress the clutter and speckle noise. This is beneficial for the later ability of the deep neural network to extract accurate and discriminative semantic information about the aircraft. Further, an improved Faster R-CNN is developed with a novel pyramid network constructed by fusing local and contextual attention. The local attention adaptively highlights the significant objects by enhancing their distinguishable features, and the contextual attention facilitates the network to extract distinct contextual information of the image. Fusing the local and contextual attention can guarantee that the aircraft is detected as completely as possible. Extensive experiments are performed on TerraSAR-X SAR datasets for benchmark comparison. The experimental results demonstrate that the proposed aircraft detection approach could achieve up to 91.7% of average precision in low SCNR, showing effectiveness and superiority over a number of benchmarks.

Keywords: synthetic aperture radar (SAR); aircraft detection; deep learning; scattering enhancement; attention mechanism

check for updates

Citation: Zhang, X.; Hu, D.; Li, S.; Luo, Y.; Li, J.; Zhang, C. Aircraft Detection from Low SCNR SAR Imagery Using Coherent Scattering Enhancement and Fused Attention Pyramid. *Remote Sens.* **2023**, *15*, 4480. https://doi.org/10.3390/rs15184480

Academic Editor: Dusan Gleich

Received: 30 July 2023
Revised: 2 September 2023
Accepted: 9 September 2023
Published: 12 September 2023

1. Introduction

As an active microwave imaging sensor, synthetic aperture radar (SAR) has the characteristics of large-scale Earth observation, penetrating clouds and fogs, and all-day and all-weather data acquisition. SAR images have been widely applied in many fields, such as environmental protection, ocean monitoring, and military domains [1,2]. Target detection is a typical application of SAR images, attracting a number of research in academia and industry [3]. Amongst them, aircraft detection is a major task that plays an important role in airport management as well as battlefield reconnaissance.

Before the emergence of deep learning, there were three main categories in traditional SAR-based target detection algorithms, including target structure and geometric features, texture features, and statistical analysis. For a specific target, its structural characteristics or geometry can provide important information prior to target detection. For an aircraft, its structure is usually a "Y" or "T" shape. Gao et al. [4] proposed a target interpretation method based on aircraft geometric features for high-resolution SAR images. The Hough transform was used to extract the skeleton composed of the wings and the fuselage, and the other parts of aircraft were identified based on the collinearity of aircraft structure and symmetry. Guo et al. [5] used edge detection algorithms based on the Canny operator to extract the candidate slices of the aircraft target. Textures are another set of key features commonly extracted from SAR images to describe visual properties using directional gradient distribution and visual saliency [6]. Tan et al. [7] developed a gradient texture saliency mapping method based on the local gradient distribution of directions to perform aircraft detection. Li et al. [8] proposed a target detection algorithm to address the challenge in selecting a suitable SAR clutter statistical model based on double-domain sparse reconstruction saliency. He et al. [9] proposed a multi-component model based on mixed statistical distribution, integrating both the target structure information and statistical distribution. However, these traditional feature extraction methods have a limited ability to excavate high-level semantic information of SAR images.

Recently, with the rapid development of deep learning, the convolutional neural network (CNN) has shown strong capabilities for feature representation. Unlike traditional feature engineering, deep semantic features are learned from CNNs with superior discriminative and generalized ability. There are a number of popular CNN architectures, such as AlexNet [10], GoogleNet [11], and ResNet [12]. Target detection, as one of the important tasks in computer vision, has benefited remarkably from those deep networks. Many studies have shown excellent object detection results using optical images from different deep networks, including Cascade R-CNN, Faster R-CNN, YOLO series.

SAR images are different from optical ones, however. CNNs have also demonstrated superior performance on SAR-based target detection through feature learning and feature representation. For example, Li et al. [13] proposed an improved Faster R-CNN for ship detection in SAR images. Jiao et al. [14] used densely connected feature maps from different network layers to solve the issue of multi-scale and multi-scene SAR-based ship detection. Cui et al. [15] connected the attention module with the pyramid to acquire abundant features. Li et al. [16] excavated complementary features in spatial and frequency domains to boost detection accuracy. To detect densely arranged targets, Yang et al. [17] employed rotated bounding box to detect ships using SAR images.

Although existing deep learning-based detectors have achieved promising results, there are still open challenges for using SAR images to detect aircraft. First, SAR images are inevitably contaminated by speckle noise as the active sensing system receives varying degrees of microwave signals. Second, land clutter in complicated scenes generally has similar scattering intensity as targets, causing interference with target detection. These compound effects result in a low signal-clutter-noise-ratio (SCNR) circumstance for SAR-based aircraft detection. Deep learning-based detectors which use such low SCNR SAR aircraft images as input would naturally fail to learn discriminative features, leading to false or missing alarms. Third, an aircraft usually appears as a set of discrete scattering centers in SAR image, which is different from how it appears in optical images. Further, the commonly used feature pyramid network (FPN) in deep learning ignores the target's contextual information, resulting in deterioration of detection performance. FPN also utilizes upsampling to concatenate information from different layers, which leads to serious loss of high-level semantic information during the upsampling process. Figure 1 illustrates aircraft in the optical images and SAR images with low SCNR. All the above factors have influenced on the performance of current deep learning-based aircraft detection algorithms, which is an extremely challenging task.

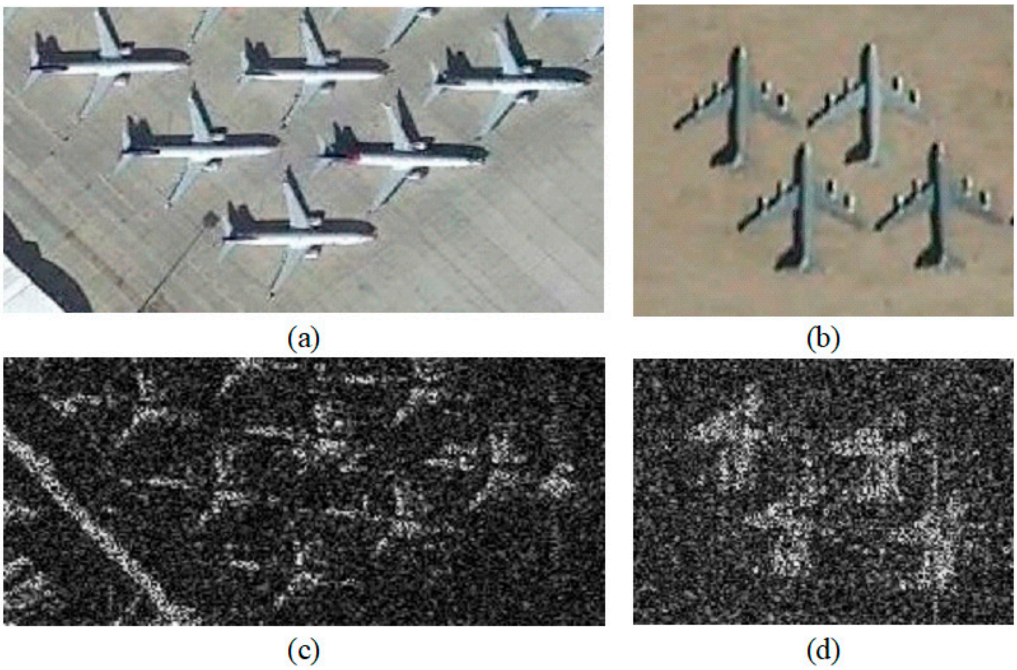


Figure 1. Aircraft in optical images and SAR images with low SCNR. (**a**,**b**) Google Earth images. (**c**,**d**) SAR images from TerraSAR-X corresponding to (**a**,**b**), respectively.

To address the problems outlined above, we propose a novel aircraft detection framework using low SCNR SAR imagery that is based on coherent scattering enhancement (CSE) in single-look complex (SLC) data and a modified Faster R-CNN with Fusion Local and Contextual Attention Pyramid Network (FLCAPN). There are large scattering characteristics discrepancies between the human-made targets and natural background; targets have strong coherence within a large angular mismatch range, while the background requires sub-pixel registration to form strong coherence. Based on these principles, we designed the CSE preprocessing to enhance aircraft information in low SCNR SAR images before feeding them into the detector. The FLCAPN is embedded in the basic Faster R-CNN to reduce the false alarms and to enhance the localization capability. Within FLCAPN, the local attention module realizes adaptive local attention to target features, thereby reducing the negative impact from clutter and speckle noise. Features are then fused with the upper layer features and sent to the contextual attention module to capture contextual information from a full feature map. In doing so, the fusion attention mechanism in FLCAPN can provide more distinguishable semantic features, achieving accurate target location.

The main contributions of this paper are summarized as three folds:

(1) For aircraft detection in low SCNR SAR images, the CSE is introduced and integrated to construct the Faster R-CNN-based detector. The CSE preprocessing can apparently enhance the scattering information of the aircraft and reduce the background clutter and speckle noise.

(2) We propose a novel FLCAPN attention pyramid that aggregates the features with local information and contextual information. In FLCAPN, the local attention can learn target local features adaptively, and the contextual attention facilitates the network in extracting significant context information from the whole image, reducing false alarms in an efficient and effective way.

(3) We construct a low SCNR SAR image dataset for aircraft detection and conduct extensive experiments via benchmark comparison. The results demonstrate the effectiveness and superiority of the proposed approach.

The rest of the paper is organized as follows. Section 2 reviews the literature that is related to this paper; Section 3 introduces the proposed aircraft detection method in detail, Section 4 shows the experimental results and analyses; Section 5 discusses the effectiveness of our method. Finally, Section 6 draws conclusions accordingly.

2. Related Work

2.1. CNN-Based Object Detection Methods

The existing deep learning object detection algorithms are primarily within two categories, including two-stage detectors and one-stage detectors. The two-stage detectors first propose regions of interest in the input image and feed these regions into the network for classification and regression. Commonly used networks include Faster R-CNN [18], Cascade R-CNN [19], etc. Faster R-CNN is further optimized on the basis of Fast R-CNN [20], where the region proposal network (RPN) is proposed to replace the region proposal module of the traditional selective search method. The RPN surpasses traditional selective search in both recall and speed. In addition, the RPN shares the same backbone with the detection network Fast RCNN, while greatly reduces the inference speed. Cascade R-CNN uses a multi-stage perceptron cascade, which has the advantage of alleviating the overfitting during training and the mismatch between proposals and ground truth during inference.

The one-stage detectors have attracted significant attention due to their faster calculation speed compared with two-stage approaches. The typical one-stage networks include Single Shot MultiBox Detector (SSD) [21], RetinaNet [22], and You Only Look Once (YOLO)v8 [23]. SSD uses feature maps from different stages to detect objects of different sizes. In RetinaNet, the Focal Loss is proposed for the first time, achieving the analogous accuracy of two-stage detectors. These detectors mentioned above are all based on anchors. Recently, a large number of anchor-free detectors have been proposed, which abandon numerous anchors and detect object by key-points or dense predictions. CornerNet [24] is one of these types of detectors, and it detects the target by predicting a pair of the object's corners. FCOS [25] predicts the distance from the center point of the object to the four sides pixel by pixel to perform object detection.

2.2. Feature Pyramid Networks in Object Detection

The backbone network generates multi-scale feature maps when extracting target features. If only a single feature map is used for detection, it cannot characterize objects across multiple scales. Therefore, the Feature Pyramid Network (FPN) [26] is proposed to handle multi-scale detection information in detection; it fuses low-level features with more spatial information and high-level features with rich semantic information through horizontal connections and top-down operation. Unlike FPN, PANet [27] adds another bottom-up path to form bidirectional connections between pyramids, showing superior performance. Compared with PANet, BiFPN in EfficientDet [28] uses repeated stacking of multiple weighted feature fusion blocks to obtain increased detection accuracy. In DetectoRS [29], Recursive Feature Pyramid (RFP) integrates the feedback from FPN and connects to the backbone network so that the features obtained by the retraining of the backbone network are suitable for detection or segmentation tasks. Recently, the neural architecture search technique has been adopted to achieve the optimal FPN structure, such as NAS-FPN [30] and Auto-FPN [31].

2.3. CNN-Based Object Detection in SAR Images

Deep learning-based target detection of SAR images has been developed rapidly based on CNN detectors. Further, current state-of-the-art SAR target detection methods focus mainly on attention mechanism. For example, Lin et al. [32] used squeeze-and-excitation attention mechanism-based Faster R-CNN. Cui et al. [33] proposed a CenterNet-based ship detection method for large-scene SAR images and designed a spatial shuffle-group enhancement (SSE) attention module to extract stronger semantic features while suppressing noise and inland interference. Fu et al. [34] designed an anchor-free feature balancing network that used an attention feature balancing pyramid and feature refinement to balance features at different levels.

As for aircraft detection, Zhao et al. [35] developed atrous convolution to expand the receptive field range and used attention modules to enhance the extraction of aircraft information. Guo et al. [36] presented a hybrid approach by combining an attention pyramid network and scattering information enhancement. A convolutional block attention module (CBAM) [37] was used to help the network focus on the aircraft target feature and avoid clutter interference. Kang et al. [38] developed a scattering point relationship module to complete the analysis and correlation of scattering points, such that the integrity of the aircraft detection was ensured. The contextual feature attention was presented to capture the global spatial and semantic information with a large receptive field, increasing the localization accuracy. Zhao et al. [39] proposed the attentional feature refinement and alignment network by fusing the attention feature module and using deformable convolution and refined predicting box. In [40], Wang et al. constructed semantic condition constraints as well as a global coordinate attention mechanism, to improve aircraft localization and recognition accuracy. A geospatial transformer framework was designed to detect aircraft in large-scale SAR images that mainly consisted of a multiscale geospatial contextual attention network [41]. To address significant intraclass differences and inconspicuous interclass variations, a global instance contrast network (GICN is proposed to improve interclass divergences and intraclass compactness [42].

However, the above studies are based on SAR amplitude images, and the phase information that is available in SAR SLC data has not been extracted. Specifically, SAR image target detection in low SCNR environments are rarely considered or investigated.

3. Methodology

The framework of the proposed SAR image aircraft detection is shown in Figure 2, which mainly consists of four parts. The first part is the preprocessing, where SAR images are input via CSE to provide high SCNR images for the subsequent network. The second part is the backbone network, which is used for semantic feature extraction. As the third part, the FLCAPN is developed by fusing the local and contextual attention mechanism, aiming to export the refined and discriminative features for the successive detector. To implement the final regression and classification, the last part of the detector is designed as a basic Faster R-CNN. Each part is described in detail below.

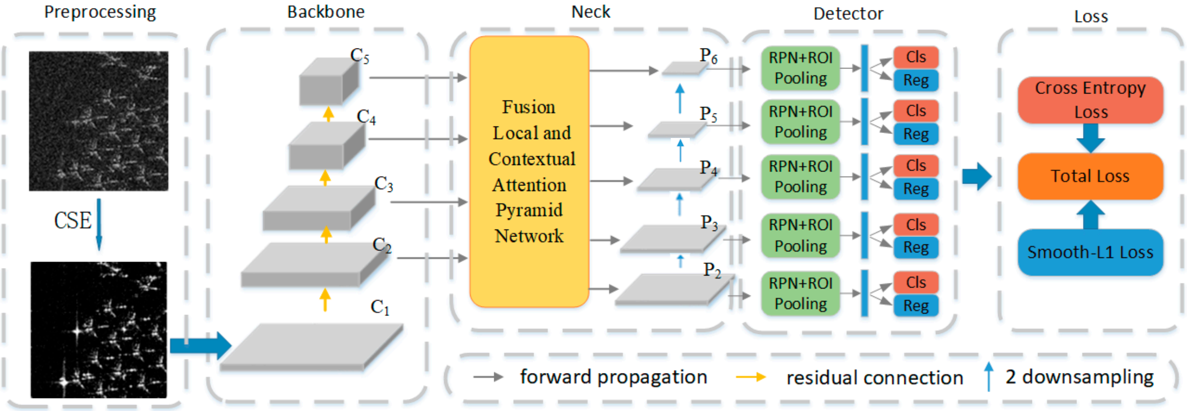


Figure 2. The overall framework of the proposed method.

3.1. CSE Preprocessing

As shown in Figure 1c,d, strong background clutter and severe speckle noise result in low SCNR SAR images, which weakens the feature information of the aircraft target. This makes it extremely difficult for the network to extract accurate aircraft target semantic features. Natural background requires accurate registration to generate coherence, whereas artificial targets show strong coherence over a wide range of angular mismatches. Considering the discrepancy between the scattering characteristics of machine-made targets and natural background, we introduce the sub-aperture CSE approach to boost target scattering information and suppress interference in low SCNR SAR images [43–45]. The reason for this operation lies in how CSE could promote SAR image quality to facilitate the network in learning to accurately target semantic features.

Figure 3 shows the sub-aperture coherent process, which includes the following steps: (1) First, apply the Fourier transform on the SAR SLC data in the azimuth direction. (2) Estimate the weights, and de-weight the generated spectrum. (3) Decompose the spectrum to create the two sub-aperture images. (4) Calculate coherence of the two sub-aperture images to obtain the azimuth coherent image. (5) Then, apply a similar procedure to SAR data in the range direction, obtaining the range coherent image. (6) Finally, the final coherent image is created by performing the incoherent calculation on the two coherent images.

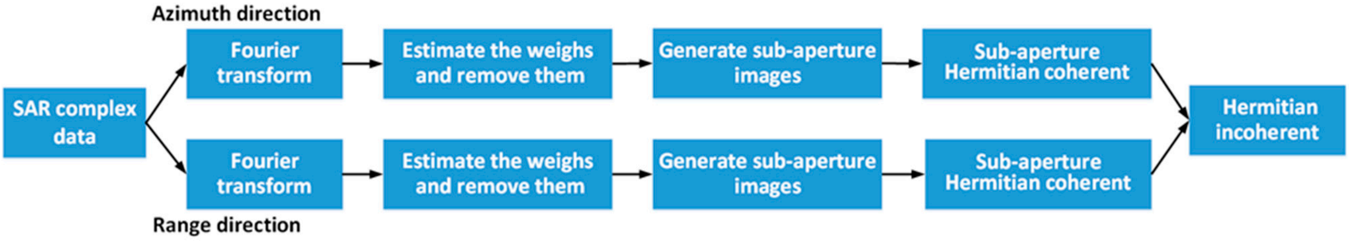

Azimuth direction
SAR complex data
Fourier transform
Estimate the weighs and remove them
Generate sub-aperture images
Sub-aperture Hermitian coherent
Hermitian incoherent
Fourier transform
Estimate the weighs and remove them
Generate sub-aperture images
Sub-aperture Hermitian coherent
Range direction

Figure 3. The overview of sub-aperture coherent process.

Given the target situated at the r_0 range of the SAR image, we apply the Fourier transform on the SAR SLC data in the azimuth direction, and the obtained frequency spectrum is as follows.

$$S(f) = \exp(\mathrm{j} \cdot \phi_0)\Pi_B(f)\exp(-\mathrm{j} \cdot 2\pi \cdot f \cdot x_0) \tag{1}$$

where $x_0 = 2r_0/c$, ϕ_0 is a constant, B is the bandwidth when $-B/2 \le f \le B/2$, $\Pi_B(f) = 1$, and $\Pi_B(f) = 0$ elsewhere.

SAR data spectra are typically convolved with a weight function to reduce the influence of the side lobe amplitude of point target impulse responses. Before generating the sub-aperture images in the azimuth direction, it is necessary to eliminate the influence of weighting via deconvolution, which is performed in two steps: (1) Estimate the weighting function in the Doppler domain by averaging the Doppler spectrum amplitude. (2) Calculate the inverse normalization function of the estimated weight function, and then apply it to the spectrum.

For simplicity, the relative motion between the target and SAR is ignored, i.e., we consider the Doppler frequency as zero. Each spectrum in azimuth direction is further separated into two halves, S_1, S_2, with a spectral width of B/2.

$$\begin{cases} S_1(f) = \exp(\mathrm{j} \cdot \phi_0) \cdot \Pi_{B/2}(f + \frac{B}{4})\exp(-\mathrm{j} \cdot 2\pi \cdot f \cdot x_0) \\ S_2(f) = \exp(\mathrm{j} \cdot \phi_0) \cdot \Pi_{B/2}(f - \frac{B}{4})\exp(-\mathrm{j} \cdot 2\pi \cdot f \cdot x_0) \end{cases} \tag{2}$$

where S_1, S_2 are two sub-aperture images, and their equivalents in the spatial domain are written as follows.

$$\begin{cases} s_1(t) = \exp(\mathrm{j}\phi_0) \cdot \sin c(\frac{\pi \cdot B(x - x_0)}{2}) \cdot \exp(\frac{-\mathrm{j} \cdot \pi \cdot B \cdot (x - x_0)}{2}) \\ s_2(t) = \exp(\mathrm{j}\phi_0) \cdot \sin c(\frac{\pi \cdot B(x - x_0)}{2}) \cdot \exp(\frac{\mathrm{j} \cdot \pi \cdot B \cdot (x - x_0)}{2}) \end{cases} \tag{3}$$

where s_1, s_2 are SAR SLC data in spatial domain. Afterwards, the Internal Hermitian Product (IHP) is adopted to calculate the coherence between the two sub-images, and the formula is as follows.

$$\xi_{hem} = \left\langle s_1 \cdot s_2^* \right\rangle \tag{4}$$

where $\langle \cdot \rangle$ represents the spatial neighborhood average. In contrast to other approaches that only compute the phase similarity of sub-aperture images, the IHP effectively captures both the amplitude and phase information. A similar procedure can be used to obtain the sub-aperture coherent image in range direction.

Figure 4 offers a diagram of each step in the procedure of azimuthal coherence processing. As seen in Figure 4h, the aircraft targets in the SAR image after the incoherent processing are significantly clearer and the scattering centers are more prominent, while the clutter and speckle noise are severely suppressed. Obviously, the image processed by CSE is more favorable for extracting the aircraft target features by deep neural networks.

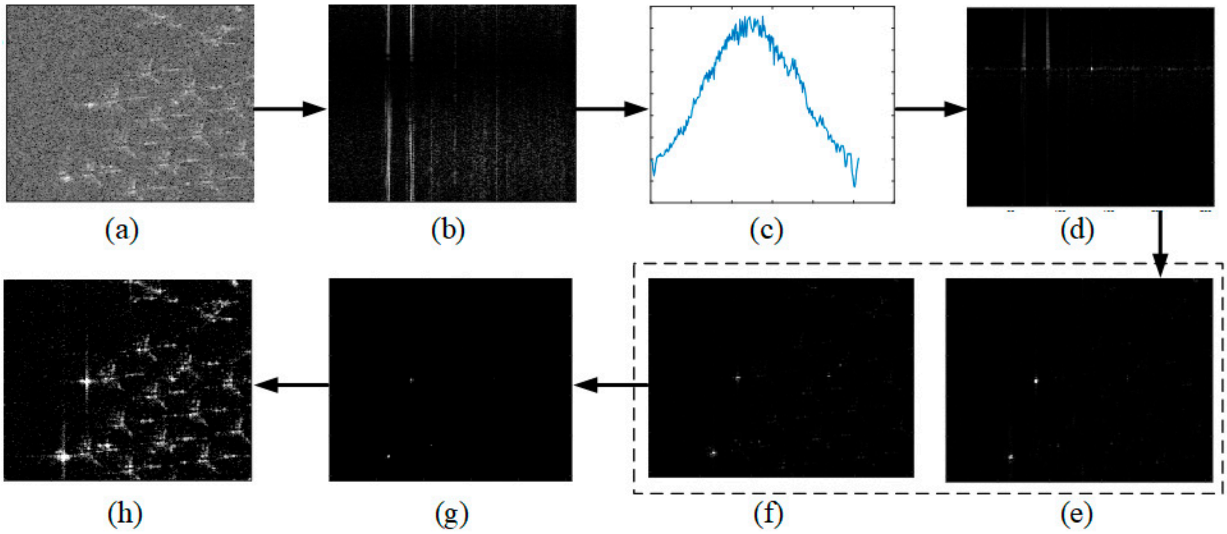


Figure 4. Diagram of all steps in the azimuthal coherence processing. (**a**) The amplitude of SAR SLC data. (**b**) Azimuth spectrum. (**c**) Estimated weight. (**d**) De-weighted spectrum. (**e**) Azimuth-right spectrum after segmentation. (**f**) Azimuth-left spectrum after segmentation. (**g**) The azimuthal coherence image. (**h**) The adjusted scattering enhancement image. (Note: In (**e**,**f**,**g**), many strong scattering points of the aircraft target are not visible because the image contrast has not been adjusted).

3.2. Fusion Local and Contextual Attention Pyramid Network

Many object detection networks use FPN to exploit multi-scale semantic information. Specifically, the FPN obtains a feature map that contains multi-scale semantic information by fusing the low-level features directly with high-level semantic information. However, due to the interference of speckle noise and clutter in the SAR image, the detection of aircraft produces numerous false alarms with inaccurate detected location. Inspired by the work in [46,47], FLCAPN was developed to organically improve the learning ability, detection accuracy, fusing local attention (LA), and contextual attention (CA) mechanism. LA is designed by following the basic structure of the bottleneck attention module in [46], except for adopting the channel shuffle operation. It can adaptively pay more attention to the characteristics of the aircraft rather than the clutter. Motivated by the crisscross attention for strong ability to learn contextual information [47], we adapted it to present CA to compensate for the lack of information surrounding the target, which is by convolution and LA. Note that the CA does not consist of a recurrent operation, which is different from the original crisscross attention. CA aims to adaptively capture contextual information on the horizontal and vertical path, improving detected box location accuracy. In summary, by integrating the complementary LA and CA, the proposed FLCAPN is able to learn local target features and contextual information effectively. The overall structure of FLCAPN is shown in Figure 5, where $C_i (i = 2, 3, 4)$ represents the features extracted from the backbone network and P_i represents the pyramid features finally output to the detector. This process is calculated as follows.

$$C_i' = Conv_{1\times1}(FA(C_i)) \qquad i = 2,3,4,5 \tag{5}$$

$$P_i = \begin{cases} CA(Upsample(P_{i+1}) + C_i') & i = 2,3,4 \\ CA(C_i') & i = 5 \\ Maxpool(P_{i-1}) & i = 6 \end{cases} \tag{6}$$

where $Conv_{1\times1}$ represents 1 × 1 convolution. In addition, P_6 is obtained by max-pooling downsampling from P_5, where the convolution kernel size is 1 and the stride is 2. The low-level features contain richer texture information, so more background clutter and speckle noise are also preserved, leading to false detections. In contrast, the high-level features contain more semantic information from the target, resulting in inaccurate target positioning. Therefore, the low-level features first go through the LA; then, the network focuses more on the target itself, thereby reducing the interference of background clutter and noise. Element-wise addition with high-level features allows the network to obtain rich semantic information. In order to make up for the neglect of the background position information around the target caused by LA, CA is leveraged to strengthen the difference information between the learning target and the surrounding background, which is beneficial for obtaining more accurate detection boxes.

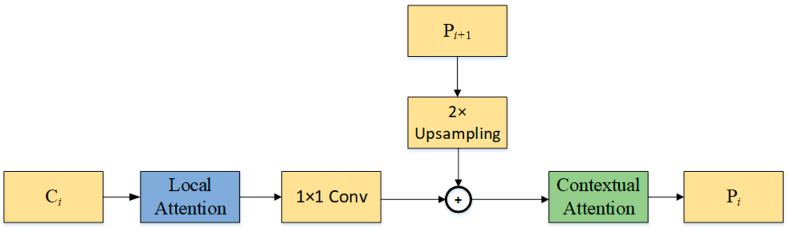


Figure 5. The overall structure of FLCAPN.

(1) Local attention: Considering the SAR imaging principle, the target image can be seen as a series of scattering centers that are difficult to detect due to the influence of speckle noise and clutter. LA is excavated to reduce the negative impact of noise and clutter so that the network can adaptively focus on aircraft targets. The overall architecture of the LA module is illustrated in Figure 6.

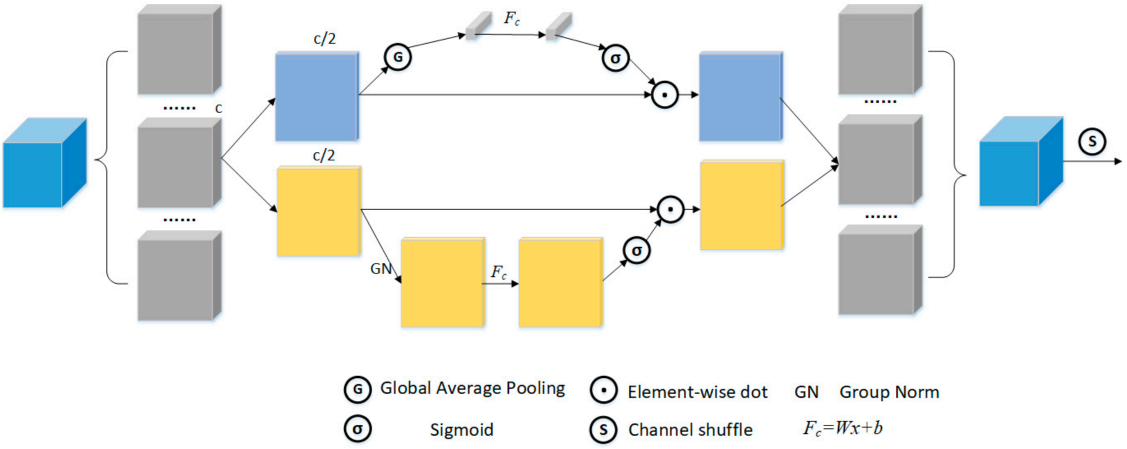


Figure 6. The architecture of LA.

Consider a feature map $X \in \mathbb{R}^{nC \times W \times H}$, where nC, W, and H represents the number of channels, width, and height, respectively. Firstly, divide X into n groups in the channel dimension to obtain sub-features $X_k \in \mathbb{R}^{C \times W \times H}$ $(k = 1, 2, \ldots, n)$. Then, the sub-features are each sent to the attention module to acquire the corresponding coefficients. Finally, all sub-features are fused to determine the final feature.

Specifically, each sub-feature is divided into two branches, $X_{k1}, X_{k2} \in \mathbb{R}^{C/2 \times W \times H}$. One obtains the connection between the channels, and the other obtains the connection between the feature spaces. Compressing X_{k1} to $s \in \mathbb{R}^{C/2 \times 1 \times 1}$ with global average pooling, the process is calculated as:

$$s = G(X_{k1}) = \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} X_{k1}(i, j) \tag{7}$$

where $G(\cdot)$ is global average pooling. The information is then aggregated using the sigmoid activation function to generate the final channel feature map:

$$X'_{k1} = \sigma(\mathrm{F}_c(s))X_{k1} = \sigma(w_1 s + b_1)X_{k1} \tag{8}$$

where $w_1 \in \mathbb{R}^{C/2 \times 1 \times 1}$ and $b_1 \in \mathbb{R}^{C/2 \times 1 \times 1}$ are parameters to scale s.

Furthermore, another branch focus on the association of features in space, i.e., on where. First, we use Group Norm (GN) for X_{k2} to determine spatial information, then use fully connected layer to enhance the feature representation of X_{k2}, and retrieve the final spatial attention:

$$X'_{k2} = \sigma(w_2 GN(X_{k2}) + b_1)X_{k2} \tag{9}$$

where $w_2 \in \mathbb{R}^{C/2 \times 1 \times 1}$ and $b_2 \in \mathbb{R}^{C/2 \times 1 \times 1}$. Then, the information of these two branches is aggregated by channel fusion:

$$X'_k = Concat(X_{k1}, X_{k2}) \tag{10}$$

where $X'_k \in \mathbb{R}^{C \times W \times H}$ $(k = 1, 2, \ldots, n)$. Afterwards, all sub-features are obtained. The LA design follows the principle of a human visual system, which has 'what' (channel) and 'where' (spatial) pathways, and both pathways contribute to processing visual information. In order to effectively fuse the information from the two attention channels and make full use of the information from all channels, similar to shuffleNetv2, the channel shuffle operation is used to fuse information across groups in the channel dimension. Finally, a feature map with the same size as the input is obtained.

(2) Contextual attention: In order to make the network capture the information around the target, the CA is designed to obtain the difference between the target and the surrounding background. It is implemented by adding upper-level features and local features obtained through LA. The process is shown in Figure 7.

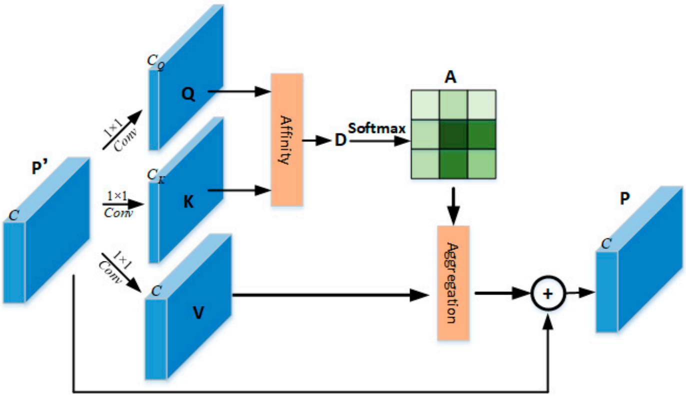


Figure 7. The architecture of CA.

Given a local feature map $P' \in \mathbb{R}^{C\times W\times H}$, firstly, perform three 1×1 convolution to generate three feature maps, $Q, K, V \in \mathbb{R}^{(C_Q/C_K/C_V)\times W\times H}$, where $C_Q = C_K < C_V = C$. Then, aiming to achieve contextual attention at position u in the spatial dimension of Q, we can acquire vector $Q_u \in \mathbb{R}^{C_Q}$, and vector $K_{iu} \in \mathbb{R}^{C_K}(i = 1, 2 \ldots W + H - 1)$, which are in the same row and column of the corresponding position u in the feature map K. Now, multiply the vectors of Q_u by the transpose of K_{iu} to form a new vector with dimension $W + H - 1$. The operation is calculated as follows:

$$d_{i,u} = Q_u K_{iu}^T \tag{11}$$

where $d_{i,u} \in \mathbb{R}^{W+H-1}$. Perform this operation on each position in Q to obtain a new feature map D with size $\mathbb{R}^{(W+H-1)\times W\times H}$. Then, SoftMax is conducted on the feature map D to acquire the normalized feature map A. Finally, we can also acquire vector set $V_{i,u} \in \mathbb{R}^{C_V}$ $(i = 1, 2 \ldots W + H - 1)$, which is in the same row and column with position u. The contextual information of position u is calculated as:

$$P_u = \sum_{i=0}^{W+H-1} A_{i,u} V_{i,u} + P'_u \tag{12}$$

where P_u, P'_u is the feature in $P, P' \in \mathbb{R}^{C\times W\times H}$ at position u and $A_{i,u}$ is a scalar value at channel i and position u in the feature map A. Perform this operation on each position to determine the contextual feature map P.

By using this CA structure, contextual information in horizontal and vertical directions can be collected to enhance pixel-wise representative capability with light-weight computation. In particular, the affinity and aggregation operations bring CA a wide contextual view and selectively aggregate contexts according to the spatial attention map. After CA, the contextual information of the image is obtained on the basis of fusing the upper layer features and LA features, which helps to distinguish aircraft targets in surrounding backgrounds and locate aircraft targets more accurately.

3.3. Loss Function

Similar to basic Faster R-CNN, the proposed network is optimized using a multi-task loss function.

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i{}^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i{}^* L_{reg}(t_i, t_i{}^*) \tag{13}$$

where p_i represents the probability that the i-th Anchor is predicted to be the true label; $p_i{}^*$ is 1 when the anchor is a positive sample and is 0 when the anchor is a negative sample; t_i represents the bounding box regression parameter used to predict the i-th Anchor; $t_i{}^*$ represents the true box label corresponding to the i-th Anchor; L_{cls} is the classification loss function, using cross entropy loss; N_{cls} is the number of anchor boxes for the classification; L_{reg} is the regression loss using Smooth L_1 loss; N_{reg} is the size of feature map; λ is the balance factor. Classification loss is defined as:

$$L_{cls}(p_i, p_i{}^*) = -\log[p_i{}^* \log(p_i) + (1 - p_i{}^*) \log(1 - p_i)] \tag{14}$$

Regression loss is defined as:

$$L_{reg}(t_i, t_i{}^*) = smooth_{L_1}\left(t_i - t_i{}^*\right) \tag{15}$$

where Smooth L_1 loss is defined as:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{16}$$

4. Experiments and Analysis

4.1. Dataset and Setting

There are no publicly available datasets for SAR aircraft target detection in low SCNR environment. Here, we built a dataset for this study to evaluate the proposed approach (Table 1). The raw SAR data of this dataset were acquired from two large scenes using the Terra SAR-X satellite. The spatial resolution of the TerraSAR-X data from the two scenes is 1 m × 1 m, and the Google Earth optical images corresponding to the scenes are shown in Figure 8. These two TerraSAR-X images contain a large number of aircraft targets located on airport runways or aprons and numerous scrapped aircraft located in the surrounding sandy environment. It should be noted that the SAR images in Figure 8 are a thumb map processed by the satellite company which has been finely filtered for speckled noise and clearly presented. However, this is not the actual detection data, which are SLC data seriously polluted by speckled noise, as shown in Figure 9. By using a manual labeling and cropping operation, 312 image patches were obtained from the two large-scene SAR images, each of which was 256 × 256. First, the 312 images were divided into a training set, a validation set, and a testing set. Since deep neural networks training requires sufficient samples, these image patches in the three sets were augmented by rotation and mirror symmetry individually, obtaining 1872 images in total. The augmentation operation was conducted separately in different datasets, so there were no duplicate samples appearing in different datasets.

Table 1. Information of dataset.

Dataset	Scene 1	Scene 2
Resolution	1 m	1 m
Polarization	HH	HH
Size	11,132 × 6251	11,166 × 6082

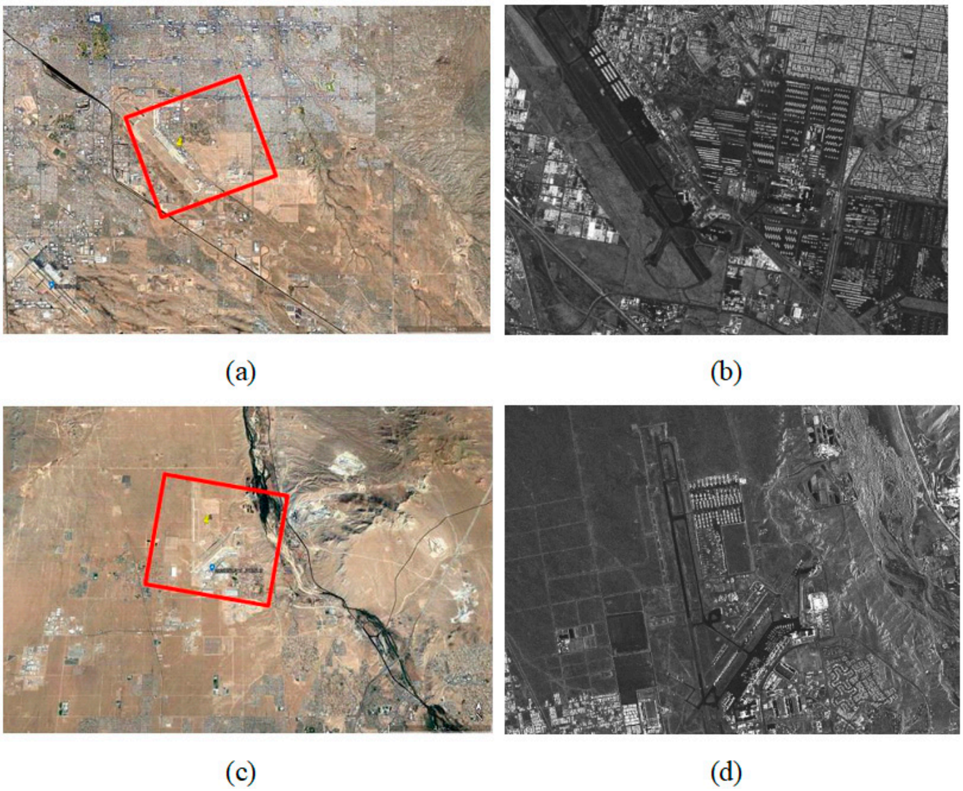


Figure 8. SAR image and corresponding optical image. (**a**,**c**) are the Google Earth optical maps. (**b**,**d**) are the SAR images corresponding to (**a**,**c**), respectively.

We implemented the experiment using Pytorch 1.7 and Cuda 11.0 and an NVIDIA GeForce RTX 3070 GPU. The backbone of the proposed method is the ResNet-50 initialized with ImageNet pretraining weights, and the dataset was randomly divided into a training set and a testing set according to the ratio of 8:2. The model was trained using the Stochastic Gradient Descent (SGD) algorithm with the learning rate set to 0.005, weight decay set to 0.0005, and momentum set to 0.9.

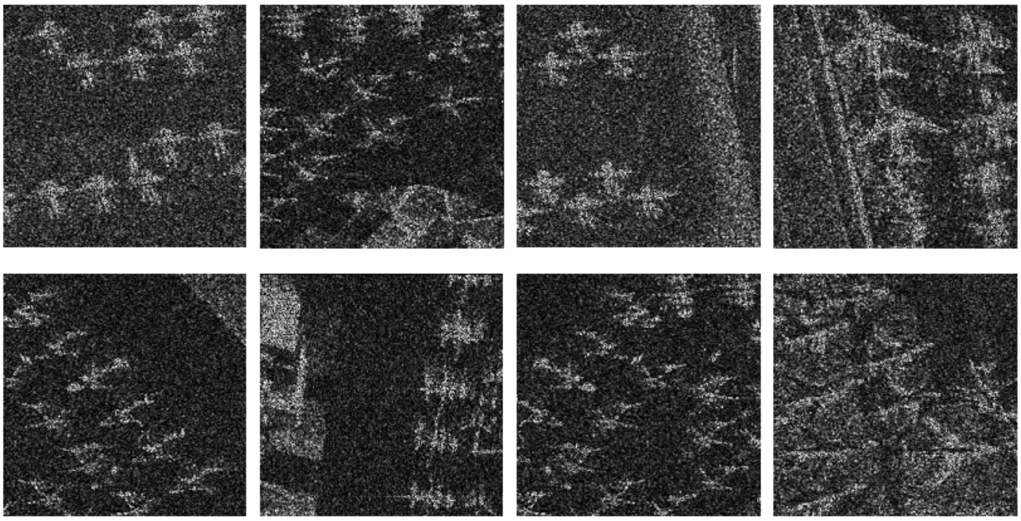


Figure 9. Sample of image slices to be detected cut from TerraSAR-X's large-scene SAR images.

4.2. Evaluation Metric

Six average precision indicators from Microsoft COCO were adopted to evaluate the performance of the aircraft detection task, including AP, AP_{50}, AP_{75}, APs, AP_{m}, and AP_{l}. Moreover, AP evaluates average precision scores by ten Intersection of Union (IoU) thresholds between predictions and ground truths (0.50:0.05:0.95). AP_{50} and AP_{75} evaluate average precision scores at 0.5 and 0.75 IoU, respectively. AP_{s}, AP_{m} and AP_{l} refer to the average precision scores of the small, medium, and large aircraft detection methods under ten IoU thresholds, the same as AP. Additionally, precision (p) refers to the proportion of all positive samples that are correctly identified as positive samples. Recall (r) is the proportion of predicted samples that are correctly identified as positive samples. The calculation of the two evaluation indicators are as follows.

$$p = \frac{TP}{TP + FP} \tag{17}$$

$$r = \frac{TP}{TP + FN} \tag{18}$$

where TP (true positives) is the number of aircraft detected correctly, FP (false positives) is the number of targets misclassified as aircraft, and FN (false negatives) is the number of aircraft misclassified as other targets.

AP, defined as the area under the precision-recall curve, is the most common metric for object detection, which is computed as:

$$AP = \int_0^1 p(r)dr \tag{19}$$

AP_{50} was chosen as the main evaluation metric, and it was mainly discussed in the following experiments.

4.3. Effect of CSE

Figure 10 shows the comparison of SAR images before and after CSE. It is obvious that the images processed with CSE are clearer than the corresponding original images. As can be observed, the land clutter around the aircraft is significantly suppressed, the target scattering centers are more prominent, and the aircraft outlines are clearer. These results illustrate that CSE can effectively improve the quality of low SCNR SAR images, and greatly suppress incoherent background clutter and speckle noise. Therefore, the processed CSE highlights the contour and scattering center features of the aircraft target, making it conducive to the discriminative semantic features extraction of the deep neural network. Furthermore, the superiority of CSE is also validated in the ablation experiments depicted in "E. Ablation Studies".

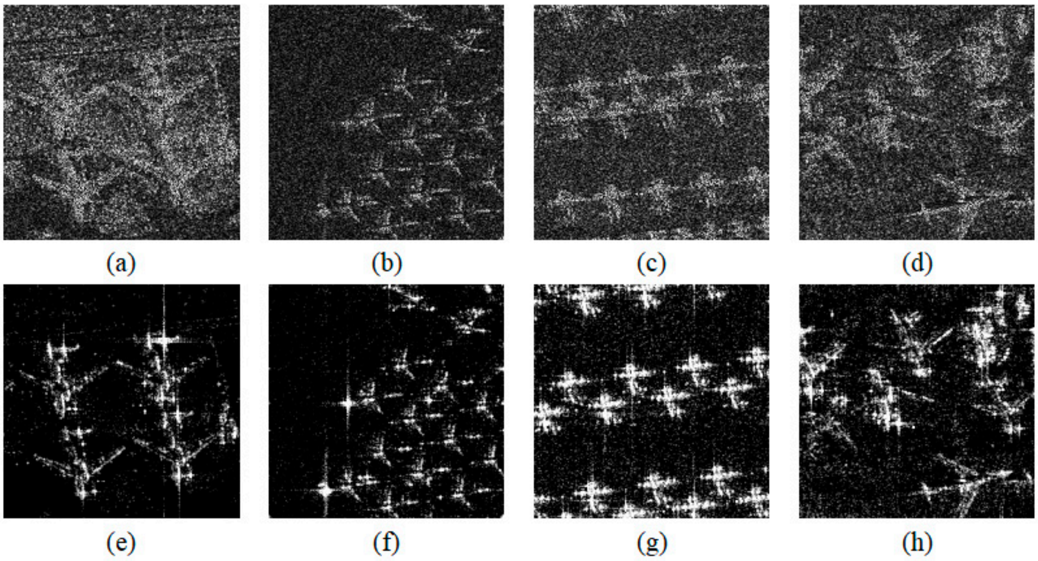


Figure 10. Comparison of low SCNR SAR images before and after CSE. (**a**–**c**) and (**d**) are the original SAR images, (**e**–**g**) and (**h**) are SAR image after CES processing of **a**–**d**, respectively.

4.4. Effect of FLCAPN

Grad-CAM [48] was used to visualize the network feature maps to effectively demonstrate the role of FLCAPN. As shown in Figure 11, the first column is the input test SAR image, the second column is the heatmap of FPN, and the third column is the heatmap of FLCAPN. All heatmaps were extracted from the features before being fed into the detector. It can be clearly observed that, compared with FPN, FLCAPN can extract the semantic information of aircraft targets more precisely and effectively, while ignoring the influence of clutter.

Figure 12 shows the superiority of our attention mechanism, where correctly detected aircraft, missed aircraft, and false alarms were highlighted by green boxes, yellow ellipses, and red circles, respectively. In the detection images by FPN of Figure 12a, we can see that the top row image misses two aircraft targets due to strong background clutter, and the bottom row image has a false alarm for land clutter. In contrast, Figure 12b shows the aircraft targets correctly detected by FLCAPN, although an aircraft target was still missed in the top row image.

Table 2 lists the detection results of FPN with LA, CA, and FLCAPN with the aim of comparing the three attention mechanisms. LA and CA increase AP_{50} by 0.6% and 0.5%, respectively, and the fusion attention mechanism increases by 1.0%, achieving better results. The role of LA is to make the network pay more attention to the important parts of the image to reduce the interference of background clutter and speckle noise. However, LA causes the network to lose the connection between the target and background, ignoring the contextual information of the targets. At this point, CA could learn the contextual representation by aggregated contextual information in horizontal and vertical directions. Therefore, the fusion attention pyramid obtains the best detection ability.

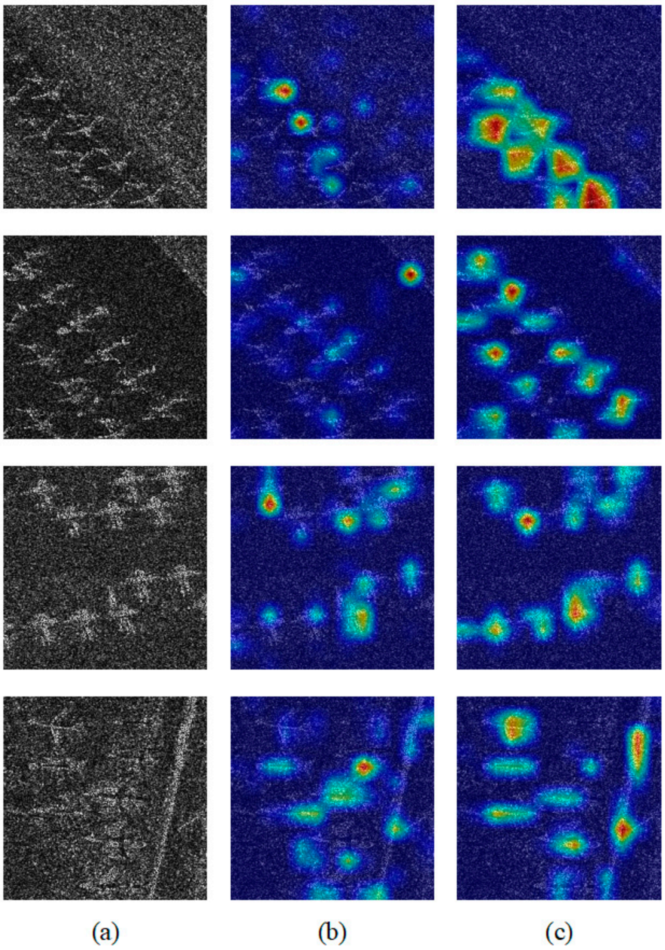


Figure 11. Feature map visualization for FPN and FLCAPN. (**a**) input images, (**b**) FPN, (**c**) FLCAPN.

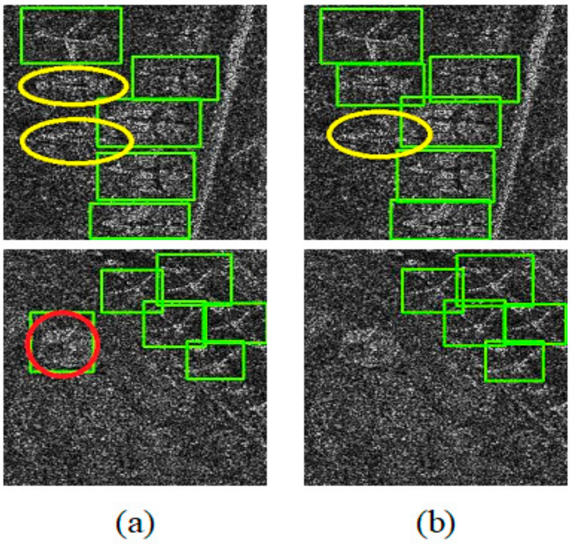


Figure 12. Samples of detection results of the FPN and FLCAPN. The green boxes and the yellow and red ellipses represent detected results, missed alarms, and false alarms, respectively. (**a**) FPN. (**b**) FLCAPN.

Table 2. Effectiveness of CA and LA. Best results are in bold.

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FPN	0.503	0.891	0.519	0.426	0.567	**0.723**
FPN + LA	0.505	0.897	0.524	**0.436**	0.575	0.707
FPN + CA	0.507	0.896	0.525	0.425	0.556	0.721
FLCAPN	**0.514**	**0.901**	**0.531**	0.406	**0.585**	0.711

4.5. Ablation Studies

In this section, we conducted a series of ablation experiments to test the effectiveness of each proposed module. The results of the ablation experiments are shown in Table 3. Comparing the results in the second and third row of Table 3, it can be observed that only the CSE preprocessing is used without FLCAPN, leading to the increment of 1.6% mAP. This result shows that CSE preprocessing can effectively improve network detection performance because the low SCNR SAR images in the experiments had been seriously contaminated by strong land clutter and speckle noise. The CSE can noticeably suppress clutter and noise, enhancing the target scattering information and thereby improving the quality of input data. It can also be found that only FLCAPN could boost mAP by 1.0%. It is apparent that each module in the proposed approach is beneficial to improving the detection performance, and that the combined adoption of the two modules increased mAP by 2.6% in total.

Table 3. Ablation studies of CSE And FLCAPN. Best results are in bold.

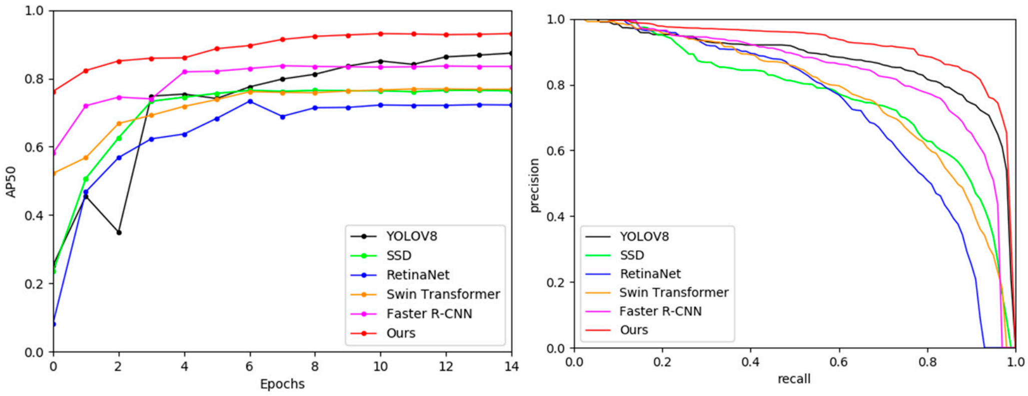
CSE	FLCAPN	AP	AP_{50}(mAP)	AP_{75}	AP_s	AP_m	AP_l
-	-	0.503	0.891	0.519	**0.426**	0.567	**0.723**
✓	-	0.519	0.907	0.528	0.406	0.570	0.677
-	✓	0.514	0.901	0.531	0.406	0.585	0.711
✓	✓	**0.534**	**0.917**	**0.561**	0.418	**0.590**	0.714

4.6. Comparison with Other CNN-Based Methods

We further conducted benchmark comparison with other CNN-based object detection networks, including the Faster R-CNN, RetinaNet, SSD-300, Swin Transformer [49], and YOLOv8, as shown in Table 4. It can be observed that, compared with other methods, the proposed approach achieved AP50 accuracy of 91.7%, which was the best detection performance. As for the AP_{75}, the proposed approach achieved an increase of 4.2% more than Faster R-CNN. From the improvement of AP_{75}, it is proven that the proposed detection framework is more accurate. This is because the background clutter and speckle noise are overwhelmingly suppressed by CSE, and the learning of important parts and the connection between the target and the background are strengthened by the fusion attention mechanism. The AP_{50} and PR curves for different methods are illustrated in Figure 13, which can further demonstrate the effectiveness of our method.

Table 4. Comparison with other methods. Best results are in bold.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN	0.503	0.835	0.519	**0.426**	0.567	**0.723**
RetinaNet	0.480	0.723	0.449	0.388	0.517	0.717
YOLOv8	0.388	0.874	0.320	0.301	0.450	0.460
SSD-300	0.465	0.764	0.453	0.367	0.503	0.643
Swin Transformer	0.378	0.768	0.315	0.332	0.417	0.367
Ours	**0.534**	**0.917**	**0.561**	0.418	**0.590**	0.714



(**a**) (**b**)

Figure 13. Comparison with other methods. (**a**) AP50 curves, (**b**) PR curves.

As shown in Figure 14, for scene (I), where large-scale aircraft are arranged close together in low SCNR SAR images, other methods have a few missed aircraft targets compared with our approach because of the disturbance caused by strong clutter. However, the proposed approach could avoid missed detections thanks to the adoption of CSE and FLCAPN. Scene (II) occurs under the complicated background of an apron building. Both Faster R-CNN and RetinaNet falsely detected the buildings as aircraft targets. SSD and YOLOv8 missed an aircraft on the edge of the scene. Our method correctly detected all aircraft without false or missed alarms. For scenes (III) and (IV), the small aircraft are densely arranged. In scene (III), RetinaNet, SSD, Swin Transformer, and YOLOv8 all have missed detections. In scene (IV), Faster R-CNN, SSD, Swin Transformer, and YOLOv8 have missed and false alarms. Additionally, RetinaNet has a case of inaccurate detection box position. As for the approach proposed in this paper, it performed significantly better than other networks, demonstrating superior adaptative ability to various scenarios.

4.7. Parameter Quantity and FPS

We also conduct experiments on the detection time and model parameter quantity (PQ) and compare the results with other methods. The experimental results are shown in Table 5. The frames per second (FPS) are used to evaluate the detection time performance. As can be seen from Table 5, the FPS of all methods are higher than those cases that use optical images as input since the SAR data fed into all networks are grayscale images. It is obvious that YOLOv8 has the highest FPS among all methods as its PQ is the lowest. The proposed approach uses fused attention mechanisms, so its PQ is the highest, leading to a relatively lower FPS. However, compared with the Swin Transformer, although our method has more parameters than the former, the calculation speed for our method is still faster. In fact, our method pays more attention to the detection accuracy and is less concerned about detection efficiency. Determining how to promote detection efficiency in the proposed approach will be our next research task.

Table 5. The parameter quantity (PQ) and FPS comparison with other methods.

Method	Faster R-CNN	RetinaNet	YOLOv8	SSD-300	Swin Transformer	Ours
PQ	41.348 M	36.33 M	3.2 M	23.746 M	44.75 M	46.272 M
FPS	396	452	1010	243	137	310

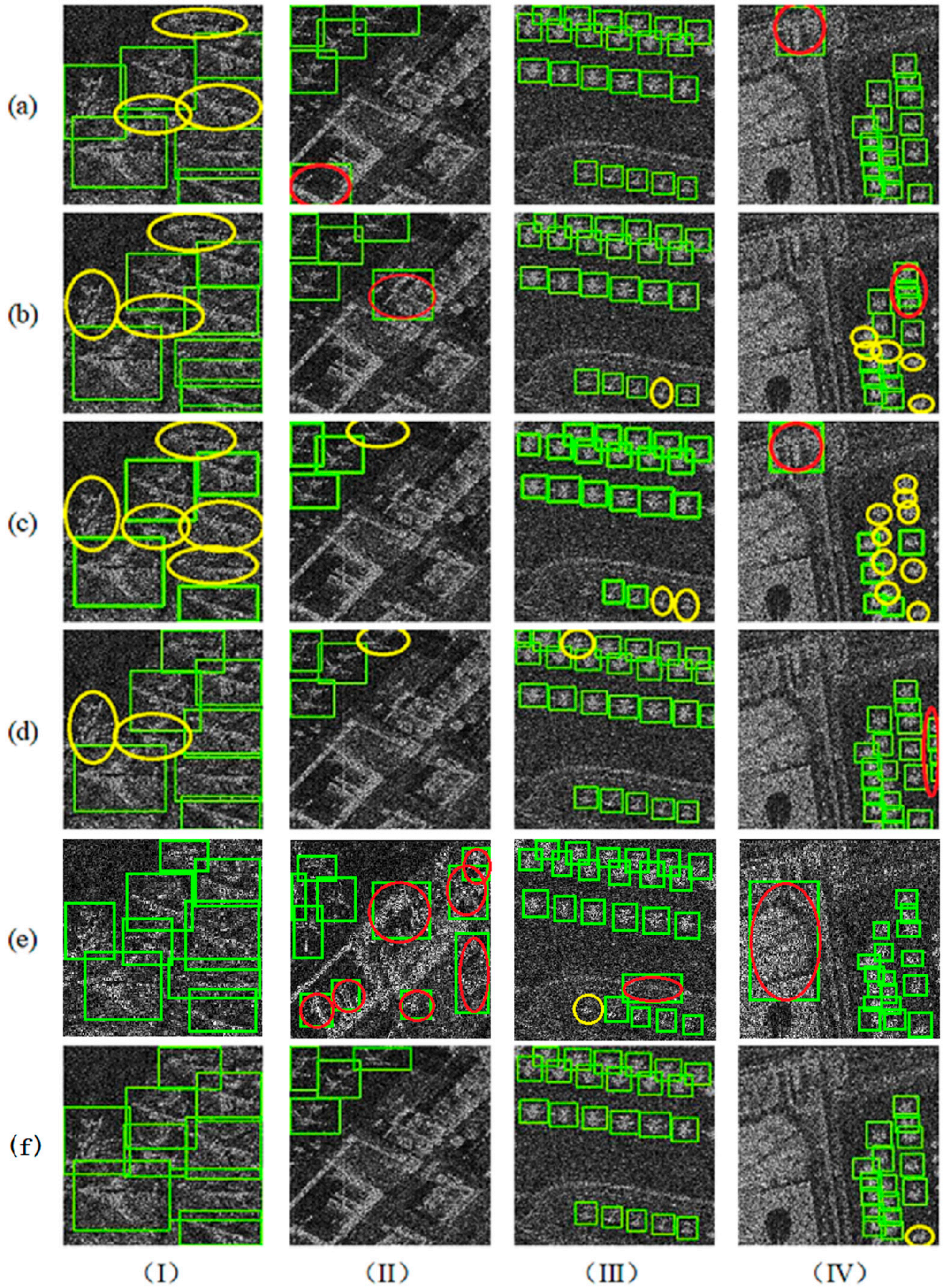


Figure 14. The detection results of different network models. The green boxes and yellow and red ellipses represent detected results, missed alarms, and false alarms, respectively. (**a**) Faster R-CNN, (**b**) RetinaNet, (**c**) SSD-300, (**d**) YOLOv8, (**e**) Swin_Transformer, and (**f**) Our approach. In addition, each column rep resents a single scene, and the four scenes are represented by (**I**), (**II**), (**III**), and (**IV**).

5. Discussion

Existing SAR aircraft detection research pays little attention to the case of low SCNR environments. This paper concentrates on aircraft detection under low SCNR. In low SCNR SAR images, it is difficult for the network to effectively learn target features due to the interference of strong clutter and speckle noise, resulting in a large number of false or missed alarms. In this paper, CSE processing was used to transform the low SCNR SAR image into a clean SAR image that is conducive to teaching the network about aircraft target features. Different from the general methods that only use SAR amplitude images, CSE makes full use of the amplitude and phase information in SAR SLC data. The principle of the CSE is based on the scattering mechanism discrepancy between artificial targets and land clutter or speckle noise. In this way, sub-aperture coherent processing in CSE is leveraged to effectively suppress land clutter and speckle noise, thereby improving the target scattering response. Therefore, CSE could facilitate the representation of the network, enhancing the aircraft target detection capability under low SCNR.

FLCAPN is proposed to enhance the ability to extract aircraft target features by fusing LA and CA, whereby the network can adaptively focus on the features of the aircraft target and capture it as discrete scattering points in the SAR image. Specifically, LA is used to refine features intelligently through channels and spatial branches. For each position, the CA aggregates contextual information in its horizontal and vertical directions to compensate for the lack of global information caused by LA and convolution operations. Overall, FLCAPN highlights features of the target, guarantees the aircraft detection integrity, and improves the accuracy of the detection box.

Although this paper studies the aircraft target detection, the proposed method can be extended to ships or any other target detection tasks using low SCNR SAR images. In addition, the CSE can also be used for SAR target recognition pretreatment. These topics will be research directions for our follow-up work.

6. Conclusions

In this paper, an aircraft detection method was designed for low SCNR SAR images in the complicated scenes. The proposed method is based on Faster R-CNN framework, integrating CSE preprocessing and FLCAPN. By introducing the CSE, the input low SCNR SAR image is transformed to a clean SAR image. Thus, the aircraft target scattering information is effectively enhanced, and the land clutter and speckle noise are well inhibited, thereby avoiding a large number of false alarms. FLCAPN was developed to aggregate the semantic information of different layers by organically fusing local information and contextual information. In FLCAPN, LA is presented to dramatically focus on the meaningful target features rather than clutter. CA enables the network to learn contextual information, which helps it learn the correlation between the scattering points of an aircraft target and the difference between the target and the surrounding background. The experimental results demonstrate that our method is generally beneficial to aircraft detection in low SCNR SAR images. In addition, both CSE and FLCAPN can be extended to other SAR target detection or recognition tasks.

Author Contributions: Software and Validation, D.H.; Conceptualization and Methodology, X.Z. and D.H.; Experiments, D.H. and Y.L.; Writing—original draft, X.Z. and D.H.; Writing—review and editing, X.Z. and J.L.; Visualization, S.L.; Supervision, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 61301224) and the Natural Science Foundation of Chongqing (Grant No. cstc2021jcyj-msxmX0174). Ce Zhang was supported in part by the Natural Environment Research Council (Grant No. 491 NE/T004002/1).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank S. Ren for providing the Faster R-CNN codes at: https://github.com/ShaoqingRen/faster_rcnn: Faster R-CNN (accessed on 15 September 2021).

Conflicts of Interest: The authors declare that they have no conflict of interest to disclose.

References

1. Wang, R.; Wang, Z.; Xia, K.; Zou, H.; Li, J. Target recognition in single-channel SAR images based on the complex-valued convolutional neural network with data augmentation. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 796–804. [CrossRef]
2. Ai, J.; Pei, Z.; Yao, B.; Wang, Z.; Xing, M. AIS data aided rayleigh cfar ship detection algorithm of multiple target environment in sar images. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 1266–1282. [CrossRef]

3. Ge, B.; An, D.; Chen, L.; Wang, W.; Feng, D.; Zhou, Z. Ground moving target detection and trajectory reconstruction methods for multichannel airborne circular SAR. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 2900–2915. [CrossRef]
4. Gao, J.; Gao, X.; Sun, X. Geometrical features-based method for aircraft target interpretation in high-resolution SAR images. *Foreign Electron. Meas. Technol.* **2022**, *34*, 21–28. [CrossRef]
5. Guo, Q.; Wang, H.; Xu, F. Aircraft target detection from spaceborne synthetic aperture radar image. *Aerosp. Shanghai* **2018**, *35*, 57–64.
6. Guo, Q.; Wang, H.; Xu, F. Research progress on aircraft detection and recognition in SAR imagery. *J. Radars* **2020**, *9*, 497–513. [CrossRef]
7. Tan, Y.; Li, Q.; Li, Y.; Tian, J. Aircraft detection in high-resolution SAR images based on a gradient textural saliency map. *Sensors* **2015**, *15*, 23071–23094. [CrossRef]
8. Li, L.; Du, L.; Wang, Z. Target detection based on dual-domain sparse reconstruction saliency in SAR images. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2018**, *11*, 4230–4243. [CrossRef]
9. He, C.; Tu, M.; Liu, X.; Xiong, D.; Liao, M. Mixture statistical distribution based multiple component model for target detection in high resolution SAR imagery. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 336. [CrossRef]
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
12. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Liu, W.; et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
13. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster r-cnn. In Proceedings of the SAR Big Data Era Models Methods Applications (BIGSARDATA), Beijing, China, 13–14 November 2017; pp. 1–6. [CrossRef]
14. Jiao, J.; Zhang, Y.; Sun, H.; Yang, X.; Gao, X.; Hong, W.; Fu, K.; Sun, X. A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection. *IEEE Access* **2018**, *6*, 20881–20892. [CrossRef]
15. Cui, Z.; Li, Q.; Cao, Z.; Liu, N. Dense attention pyramid networks for multi-scale ship detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8983–8997. [CrossRef]
16. Li, D.; Liang, Q.; Liu, H.; Liu, Q.; Liu, H.; Liao, G. A novel multidimensional domain deep learning network for SAR ship detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
17. Yang, R. A novel cnn-based detector for ship detection based on rotatable bounding box in SAR images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 1938–1958. [CrossRef]
18. 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]
19. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality 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. 6154–6162.
20. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 10–16 October 2016; pp. 21–37. [CrossRef]
22. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
23. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 17 August 2023).
24. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 765–781. [CrossRef]
25. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635.
26. 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]
27. 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]
28. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, DC, USA, 14–19 June 2020; pp. 10778–10787. [CrossRef]
29. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 10208–10219. [CrossRef]
30. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7029–7038. [CrossRef]

31. Xu, H.; Yao, L.; Li, Z.; Liang, X.; Zhang, W. Auto-fpn: Automatic network architecture adaptation for object detection beyond classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6648–6657. [CrossRef]
32. Lin, Z.; Ji, K.; Leng, X.; Kuang, G. Squeeze and excitation rank faster r-cnn for ship detection in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 751–755. [CrossRef]
33. Cui, Z.; Wang, X.; Liu, N.; Cao, Z.; Yang, J. Ship detection in large-scale SAR images via spatial shuffle-group enhance attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 379–391. [CrossRef]
34. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1331–1344. [CrossRef]
35. Zhao, Y.; Zhao, L.; Li, C.; Kuang, G. Pyramid attention dilated network for aircraft detection in SAR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 662–666. [CrossRef]
36. Guo, Q.; Wang, H.; Xu, F. Scattering enhanced attention pyramid network for aircraft detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7570–7587. [CrossRef]
37. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
38. Kang, Y. Sfr-net: Scattering feature relation network for aircraft detection in complex SAR images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
39. Zhao, Y.; Zhao, L.; Liu, Z.; Hu, D.; Kuang, G.; Liu, L. Attentional feature refinement and alignment network for aircraft detection in SAR imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
40. Chen, L.; Luo, R.; Xing, J.; Li, Z.; Yuan, Z.; Cai, X. Geospatial transformer is what you need for aircraft detection in SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
41. Wang, Z.; Xu, N.; Guo, J.; Zhang, C.; Wang, B. SCFNet: Semantic Condition Constraint Guided Feature Aware Network for Aircraft Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
42. Zhao, D.; Chen, Z.; Gao, Y.; Shi, Z. Classification Matters More: Global Instance Contrast for Fine-Grained SAR Aircraft Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
43. Souyris, J.-C.; Henry, C.; Adragna, F. On the use of complex SAR image spectral analysis for target detection: Assessment of polarimetry. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2725–2734. [CrossRef]
44. Suess, M.; Grafmueller, B.; Zahn, R. Target detection and analysis based on spectral analysis of a SAR image:a simulation approach. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Toulouse, France, 21–25 July 2003; pp. 2005–2007. [CrossRef]
45. Ferro-Famil, L.; Reigber, A.; Pottier, E.; Boerner, W.-M. Scene characterization using subaperture polarimetric SAR data. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2264–2276. [CrossRef]
46. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514. [CrossRef]
47. Huang, Z.; Wang, X.; Wei, Y.; Huang, L.; Shi, H.; Liu, W.; Huang, T.S. CCNet: Criss-Cross Attention for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
48. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradientbased localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
49. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002. [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.