SFRE-Net: Scattering Feature Relation Enhancement Network for Aircraft Detection in SAR Images

Peng Zhang, Hao Xu, Tian Tian *, Peng Gao and Jinwen Tian

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China; z_peng@hust.edu.cn (P.Z.); xu_hao_98@hust.edu.cn (H.X.); gaopengde@hust.edu.cn (P.G.); jwttian@mail.hust.edu.cn (J.T.)

* Correspondence: ttian@hust.edu.cn

Abstract: Aircraft detection in synthetic aperture radar (SAR) images is a challenging task due to the discreteness of aircraft scattering characteristics, the diversity of aircraft size, and the interference of complex backgrounds. To address these problems, we propose a novel scattering feature relation enhancement network (SFRE-Net) in this paper. Firstly, a cascade transformer block (TRsB) structure is adopted to improve the integrity of aircraft detection results by modeling the correlation between feature points. Secondly, a feature-adaptive fusion pyramid structure (FAFP) is proposed to aggregate features of different levels and scales, enable the network to autonomously extract useful semantic information, and improve the multi-scale representation ability of the network. Thirdly, a context attention-enhancement module (CAEM) is designed to improve the positioning accuracy in complex backgrounds. Considering the discreteness of scattering characteristics, the module uses a dilated convolution pyramid structure to improve the receptive field and then captures the position of the aircraft target through the coordinate attention mechanism. Experiments on the Gaofen-3 dataset demonstrate the effectiveness of SFRE-Net with a precision rate of 94.4% and a recall rate of 94.5%.

Keywords: synthetic aperture radar; aircraft detection; adaptive feature fusion; contextual coordinate attention; scattering point relation

1. Introduction

Synthetic aperture radar (SAR) has all-weather and all-day observation capability. With the unique imaging mechanism, SAR plays a crucial role in many fields [1–6], such as target detection, strategic reconnaissance, and terrain detection. Automatic target recognition (ATR) is one of the most important applications in SAR, which aims to locate and identify potential targets and has been studied for decades [7–9]. Aircraft is an important target in SAR, and the detection of it has important application value in airport management, military reconnaissance, and other fields. With the development of SAR imaging technology, SAR aircraft detection has attracted extensive attention and has become an independent research direction [10–13].

Traditional SAR image object-detection algorithms include three categories: single-feature-based, multi-feature-based, and prior-based methods. The single-feature-based methods usually use radar cross section (RCS) information to select the parts with higher pixel intensity as candidate regions. The constant false alarm rate (CFAR) [14] is the most typical single-feature-based detection method, which adopts the idea of threshold extraction and clutter statistics. On the basis of CFAR, scholars have conducted in-depth research on statistical characteristics and non-uniform backgrounds and proposed a variety of improved CFAR algorithms, such as cell-averaging constant false alarm rate (CA-CFAR) [15], variability index constant false alarm rate (VI-CFAR) [16], the dual-parameter CFAR method [17], etc. Multi-feature-based methods use a variety of features for fusion detection, such as geometric structure [18], extended fractal [19], wavelet coefficient [20], etc. The prior-based
method incorporates prior knowledge such as imaging parameters, latitude, and longitude information into the detection process to improve the detection accuracy, as performed in [21]. However, traditional SAR target detection algorithms rely on fixed features and have poor generalization ability in complex background.

Recently, deep learning has developed rapidly. The convolutional neural network (CNN) is one of the most widely used architectures in deep learning with strong feature description ability, which has made outstanding contributions in many fields [22,23]. Object-detection methods based on CNN do not need to manually design features. They have higher detection accuracy and generalization ability than traditional detection methods. Some scholars have applied deep learning methods to the field of SAR aircraft detection. Diao et al. [24] used CFAR detectors to locate potential aircraft positions and used R-CNN [25] algorithm for final detection. However, this method is not end-to-end, and the process is cumbersome. He et al. [26] designed a multi-layer parallel network structure based on components to realize the accurate positioning of aircraft targets. However, this method needs fine-grained component information, and the label-making process is complicated. An et al. [27] used a feature pyramid structure to improve the detection effect of small targets. Zhao et al. [28] built a fully convolutional attention network to extract different levels of contextual information, which improved the detection accuracy of aircraft.

Although the above-mentioned deep learning methods have made particular progress in SAR images, aircraft target detection in SAR images still faces many challenges, such as target integrity detection, target multi-scale problems, and complex background interference. First, because of the complex imaging mechanism of SAR and the differences between aircraft components, the aircraft target is discontinuous in the SAR image, which is composed of multiple discrete irregular scattering center bright spots, as shown in Figure 1a, and the key semantic information of the target is hidden between these scattering centers. The existing detection algorithms are prone to generating multiple prediction boxes for one aircraft target, as shown in Figure 1b. It is challenging to combine discrete bright spots into a whole for integrity detection. Second, since there are different aircraft models, aircraft target detection in SAR images faces the problem of size diversity, as shown in Figure 1c. Although some scholars have adopted the feature pyramid structure, such as the feature pyramid network (FPN) [29] and path aggregation network (PANet) [30], to improve the feature expression ability of the model, the above design ignores the existence of conflict information between different features, making it difficult to improve the detection accuracy further. Third, due to the irregular distribution of clutter, there are many background highlighted clutter points around aircraft, which will be confused with the aircraft parts to a certain extent, making it difficult to accurately locate and identify the target, as shown in Figure 1d.

To address these problems, we propose a scattering feature relation enhancement network (SFRE-Net). Different from the direct application of deep CNN to SAR images, SFRE-Net fully considers the discrete characteristics of aircraft targets, strengthens the structural correlation between scattering points, and suppresses clutter interference by improving the attention of scattering points. Firstly, inspired by the vision transformer [31], SFRE-Net designs a cascade transformer structure, which can construct the correlation between feature points through the self-attention mechanism and then transfer the captured correlation feature map to the scale fusion module. This method significantly enhances the structural correlation between scattering points, which is helpful for the integrity detection of aircraft. Secondly, considering that FPN [29] and other feature pyramid structures directly integrate different scales of information, which easily causes semantic conflict, SFRE-Net designs an adaptive multi-scale feature fusion structure, which controls the fusion of different scales of information through learnable weighting coefficients so that the network can independently select favorable semantic information and improve the multi-scale detection ability of the algorithm. Thirdly, SFRE-Net designs a context attention-enhancement module (CAEM). Considering the discrete characteristics of aircraft targets, CAEM uses the
dilated convolution pyramid structure to improve the receptive field of the feature map, and then adopts the coordinate attention to improve the attention of aircraft scattering points and suppress the background clutter interference. The CAEM module is beneficial to improve the positioning accuracy of aircraft targets under complex backgrounds.

![Figure 1](image.png)

**Figure 1.** Detection effect of the general detection algorithm. The green boxes, pink boxes, red boxes, and yellow ellipse represent the correct predictions, missing detections, false alarms, and complete aircrafts, respectively. (a) Aircraft targets. (b) Problems of discrete detection. (c) Aircraft targets of various sizes. (d) Problems of complex background interference.

The contributions of our work are as follows:

- The proposed SFRE-Net comprehensively considers the distribution of aircraft target scattering points and background clutter points and models the relationship between feature points, and the performance on the Gaofen-3 dataset is better than the state-of-the-art detection algorithms.
- In order to improve the integrity detection of aircraft targets, a cascade transformer block (TRsB) structure is adopted to capture the correlation between scattering feature points, effectively alleviating the phenomenon of aircraft’s discrete detection.
- In order to alleviate the semantic conflict caused by common feature pyramid structures in different scale information fusion, a feature-adaptive fusion pyramid structure (FAFP) is proposed. This design enables the network to independently select useful semantic information and improve the multi-scale expression ability of the network.
- In order to suppress the clutter interference in the background, a context attention-enhancement module (CAEM) is designed to improve the receptive field and enhance the attention of the target scattering points, which can realize the accurate positioning of aircraft targets in complex backgrounds.

The rest of the paper is arranged as follows. The related work is reviewed in Section 2. Section 3 describes the proposed aircraft-detection algorithm in detail. Experimental results, as well as performance evaluation, are presented in Section 4. Section 5 further discusses our method. Section 6 briefly summarizes this paper.
2. Related Work

This section introduces the progress of target detection algorithms in natural scenes, the development of feature pyramid structure, and the progress of SAR aircraft target detection algorithms.

2.1. Deep-Learning-Based Object-Detection Methods

Deep-learning-based object-detection algorithms have achieved great success in natural scenes and can be divided into anchor-based methods and anchor-free methods. Anchor-based detection algorithms include one-stage and two-stage methods. The two-stage detection method mainly includes two parts: region proposals extraction and detection result generation, such as R-CNN [25], fast R-CNN [32], and faster R-CNN [33]. Cascade R-CNN [34] further improves the detection accuracy by stacking cascade detectors. The two-stage detection algorithms have high accuracy but slow speed. The one-stage algorithm has higher computing speed, such as SSD [35], RetinaNet [36], and YOLOV3 [37]. However, the detection accuracy is usually lower than the two-stage algorithm. With the rapid development of one-stage algorithms, YOLOV4 [38] and YOLOV5 [39] can achieve competitive accuracy in contrast to the two-stage method.

Anchor-free detectors abandon the use of anchors. The existing anchor-free algorithms include keypoint-based methods, such as Cornernet [40] and center-based methods, such as Centernet [41] and Fcos [42]. Anchor-free detectors avoid the calculation caused by anchors and have a high recall rate. However, different from anchor-based methods, anchor-free methods lack artificial prior distribution, which leads to unstable detection results and requires more skills for re-weighting.

With the successful application of the transformer in the NLP field, the transformer has received extensive attention in computer vision, such as end-to-end object detection with transformers (DETR) [43] and the vision transformer [31]. The transformer model uses the self-attention mechanism and has global information. However, it lacks inductive bias compared with CNN, so it needs many data for training. Therefore, the combination of CNN and transformer can integrate their advantages and achieve better generalization ability on small datasets. Compared with anchor-free methods, the anchor-based methods are more stable in training and do not need many tricks to re-weight, and it is convenient to generate the required anchors through the clustering algorithm. Hence, this paper adopts the method of combining an anchor-based detector with the transformer structure.

2.2. Feature Pyramid Structure in Object-Detection Methods

Multiscale feature fusion effectively promotes the interaction of cross-scale information, which is crucial in object detection. FPN [29] is a pioneering work consisting of a top-down path and skip connections for extracting multi-scale feature representations. On the basis of FPN, PANet [30] adopts a bidirectional structure to further aggregate features of different scales. Libra R-CNN [44] balances shallow and deep features and refines the original features through the non-local block. PANet shows the effectiveness of bidirectional path fusion, but this structure struggles to deal with semantic conflicts between different scales. To further enhance the ability of multi-scale information interaction, bidirectional feature network (BiFPN) [45] designs a bidirectional path weighted fusion method.

Inspired by BiFPN, we design a feature-adaptive fusion pyramid structure (FAFP), which can automatically fuse useful information between different scales and more effectively alleviate the problem of target multi-scale detection.

2.3. SAR Aircraft-Detection Methods with Deep Learning

Unlike traditional detection algorithms, CNNs can extract features automatically, and they have better generalization ability. At present, SAR aircraft target detection based on deep CNN has attracted extensive attention. He et al. [26] designed a multi-layer parallel structure for aircraft target detection. Experiments on the TerraSAR-X dataset show that this method improved the localization accuracy of aircraft targets. An et al. [27] designed
a feature pyramid network based on rotating minimum adjacency rectangle, which is beneficial to the detection of small targets. Zhang et al. [46] adopted a cascaded three-view structure based on faster R-CNN [33], which can reduce the false alarms. Guo et al. [47] proposed an aircraft-detection algorithm based on FPN [29] structure. It combines the shallow location information and deep semantic information, improves the accuracy of detection results, and alleviates the problem of multi-scale target detection to a certain extent. In recent years, Guo et al. [11] designed a hybrid approach of scattering information enhancement and attention pyramid network to enhance the characteristics of aircraft scattering points. Based on YOLOV5s, Luo et al. [12] designed a fast and efficient two-way route aggregation attention network to achieve a balance between accuracy and speed. Kang et al. [13] designed a scatter-point correlation module to analyze and correlate scattered points. Zhao et al. [10] proposed an attentional feature refinement and alignment network to strengthen the feature-extraction ability of the model.

Previous work has made outstanding contributions to the development of aircraft target detection in SAR images. However, there are still some problems left. Firstly, the aircraft target in SAR image is discontinuous, and the existing methods find it difficult to form these discrete points into a whole for integrity detection. Second, although the existing SAR aircraft-detection algorithms use multi-scale information fusion methods, such as FPN [29], they cannot solve the problem of semantic conflict between different scales. Thirdly, because of the complex imaging mechanism of SAR, there is a problem of background clutter interference in SAR aircraft target detection, which has not been well solved. Considering these problems, we propose a scattering feature-adaptive fusion network (SFRE-Net) to solve the challenges of aircraft detection in SAR images.

3. Materials and Methods

3.1. Overview of SFRE-Net

As the latest model of the current YOLO series, YOLOV5 [39] has strong detection ability. YOLOV5s is a lightweight model in the YOLO series, achieving robust real-time detection and advanced detection accuracy on small-scale datasets. YOLOV5 uses the CSP-Darknet53 [38] as backbone, PANet [30] as Neck and YOLO detection head [48]. YOLOV5 has an excellent performance in the natural field, and we choose it as the baseline in this paper. The overall structure of SFRE-Net is shown in Figure 2.

Figure 2. The architecture of the SFRE-Net, including backbone, neck, and prediction.

SFRE-Net consists of input, backbone, neck, and prediction. The input part enriches the training dataset by mosaic [38], random horizontal flip, and random rotation data expansion methods. Inspired by YOLOV5 [39], SFRE-Net uses CSPDarknet53 [38] as the backbone, with lightweight parameters and a strong feature-extraction ability. The neck of
SFRE-Net includes a context attention-enhancement module (CAEM) and feature-adaptive fusion pyramid (FAFP) structure. FAFP adopts the transformer block (TRsB), feature-enhancement module (FEM), and adaptive spatial feature fusion (ASFF) [49] structure to improve the ability of feature aggregation. Finally, the network performs target prediction and output through the prediction. The following is a detailed explanation of our method.

3.2. Transformer Block (TRsB)

Inspired by the vision transformer [31], we adopt a transformer block (TRsB) encoder structure in SFRE-Net, as shown in Figure 3. The key idea of our method is to embed all feature points in the feature map and obtain the correlation between feature points through self-attention mechanism [50].

The typical transformer uses 1D embedding vectors as its input. To process 2D images, we reshape the input feature map $x \in \mathbb{R}^{H \times W \times C}$ into a sequence of feature points $x_p \in \mathbb{R}^{N \times C}$, where $(H, W)$ represents the original feature map resolution, $C$ denotes the channel dimension of the feature map, and $N = HW$ indicates the resulting number of feature points, which is also the effective input sequence length of the transformer encoder block. The transformer encoder block uses constant latent vector size $D$ in all its layers, so we map the feature points to $D$ dimensions with a trainable linear projection (1). Position embeddings are attached to the feature point embeddings to retain position information. In this paper, we use learnable linear 1D position embeddings. The resulting sequence of embedding vectors serves as input to the encoder; see $z_0$ in Equation (1).

The transformer encoder block is mainly composed of multi-head self-attention (MHSA) layer and multi-layer perception (MLP). Layernorm (LN) is used before every sub-layer, and residual connections are adopted after every sub-layer; see Equations (2) and (3). We also use dropout layers (Dp) in every transformer encoder block to prevent network overfitting. Equation (4) shows the final output of the transformer encoder blocks, where $L$ is the number of transformer encoders, which is set to 3 to balance the detection accuracy and detection speed in this paper.

\[
\begin{align*}
  z_0 &= x_p E + E_{pos}, \quad E \in \mathbb{R}^{C \times D}, \quad E_{pos} \in \mathbb{R}^{N \times D} \\
  z'_\ell &= Dp(MSA(LN(z_{\ell-1}))) + z_{\ell-1}, \quad \ell = 1 \ldots L \\
  z_\ell &= MLP(LN(z'_\ell)) + z'_\ell, \quad \ell = 1 \ldots L \\
  y &= Dp(z^0_L) 
\end{align*}
\]
3.3. Feature-Adaptive Fusion Pyramid (FAFP)

To fully combine the deep and shallow semantic information, we design a feature-adaptive fusion pyramid (FAFP) structure, as shown in Figure 4.

![Figure 4](image)

**Figure 4.** Description of feature pyramid structures. FPN and PANet adopt channel concatenation fusion method. BiFPN and FAFP use the weighted addition fusion method.

The FAFP adopts a two-way integration, consisting of a top-down path and a bottom-up path. Unlike other feature fusion methods, such as channel concatenation or the element-wise direct addition, the core idea of FAFP is to learn the spatial weights of different feature maps adaptively in the fusion process. The adaptive fusion process of FAFP includes two parts. Firstly, the learnable coefficients are used to guide the fusion between feature maps in the bidirectional feature pyramid. Secondly, inspired by ASFF [49], the three scale features output by the bidirectional feature pyramid are weighted aggregation to obtain the final three output scale feature maps. We represent the feature map at $l$ level ($l \in \{3, 4, 5\}$ for SFRE-Net) as $x'$, as shown in Figure 2. For $l$ level, we first scale the feature map $x^n$ of other levels ($n \neq l$) to ensure that it has the same size with $x'$, and then use the adaptive weighted addition method to obtain the final output of $l$ level. The specific formula is as follows:

$$y'_{ij} = \alpha_{ij}^l \cdot x_3^{3-l} + \beta_{ij}^l \cdot x_4^{4-l} + \gamma_{ij}^l \cdot x_5^{5-l},$$

where $x_3^{3-l}$ represents the feature vector located at $(i, j)$ in the feature map scaled from level $n$ to level $l$. $y'_{ij}$ denotes the feature vector located at $(i, j)$ in the $l$-level output feature map, $\alpha_{ij}^l$, $\beta_{ij}^l$ and $\gamma_{ij}^l$ imply the learnable weight factors in the process of $l$-level feature fusion.

To enhance the semantic richness of feature fusion, we design a feature-enhancement module (FEM) in FAFP, as shown in Figure 5. FEM improves the receptive field of the feature map through three different branches, which is beneficial to enhance the multi-scale representation capability of the model, and we adjust the kernel size according to the resolution of the feature map. FEM is used in the large-resolution feature map and FEM-s is used in the small-resolution feature map, as shown in Figure 2.

![Figure 5](image)

**Figure 5.** The structure of FEM and FEM-s. FEM is employed in the large-resolution feature map, and FEM-s is employed in the small-resolution feature map.
3.4. Context Attention-Enhancement Module (CAEM)

To improve the positioning accuracy of aircraft targets in complex backgrounds, we design a context attention-enhancement module (CAEM), as shown in Figure 6. Considering the discrete characteristics of aircraft targets in SAR images, CAEM first uses the dilated convolution pyramid to improve the receptive field of the feature map and then uses the coordinate attention mechanism [51] to strengthen the attention of target points and reduce the attention of clutter points. The coordinate attention mechanism aggregates the input features into perceptual feature maps in two independent directions, horizontal and vertical, through two 1D global pooling operations. Then, the two feature maps with specific direction information are encoded to focus on important positions. Each attention map obtains a long-distance dependency in the particular direction of the original input feature. Finally, the two attention maps with position information in a specific direction are multiplied to improve the representation of the target points of interest. In order to make the model fully extract context information, we designed an adaptive fusion method, as shown in Figure 6a. Specifically, assuming that the size of the input can be expressed as \((bs, C, H, W)\), where \(bs\) represents the batch size, we can obtain the spatial adaptive weight \([bs, 4, h, w]\) by performing convolution, concatenation, and softmax operations. Four channels correspond to four inputs, one by one. The context information can be aggregated to the output by calculating the weighted addition. CAEM fully considers the discrete characteristics of SAR image targets and effectively improves the aircraft target positioning accuracy under complex backgrounds by strengthening the attention of target points.

3.5. Prediction and Loss Function

After the adaptive feature fusion operation, the FAFP module outputs three scales of prediction feature maps with grid areas of \(80 \times 80\), \(40 \times 40\), and \(20 \times 20\), respectively. Figure 7 shows the process of target classification and prediction box regression. For each grid cell in the output feature maps, the model produces three anchors with different aspect ratios and sizes for classification and regression. The desired anchors can be obtained by clustering the size of the dataset. Then, after a \(1 \times 1\) convolution operation, each prediction box’s position, confidence, and category are predicted by classification and regression.

The irregular distribution of background clutter points in the SAR image interferes with the detection of aircraft targets, and the proportion of foreground and background in remote sensing image is uneven, resulting in the imbalance of the number of positive and
negative samples in the dataset. To alleviate these problems, focal loss [36] is used in this paper, and the formula is as follows:

$$\text{FL}(x) = \begin{cases} -a(1 - x)\gamma \log x, & y = 1 \\ -(1 - a)x\gamma \log(1 - x), & y = 0 \end{cases}$$ (6)

where $\gamma$ is set to 1.5 and $a$ is set to 0.25 in this paper.

Figure 7. The schematic diagram of prediction box generation, where purple boxes represent anchors.

4. Experimental Results

4.1. Dataset

Currently, there is only one publicly available dataset [52] in the field of SAR image aircraft detection, so we use this dataset to verify our method.

The SAR Aircraft Dataset (SAD) used in this paper is from the Gaofen-3 satellite with a resolution of 1 m. The ground truths of the SAD were discriminated and annotated by SAR ATR experts. In this paper, 3715 nonoverlapped $640 \times 640$ patches with 6556 aircraft targets are obtained by cropping the original large images. The outlines and components of aircraft targets in the dataset are clear. The size and aspect ratio distribution of aircraft targets in the SAD are shown in Figure 8. The size of different types of aircraft in the SAD varies greatly, which makes the detection difficult.

Figure 8. Distribution of aircraft target boxes in the SAR Aircraft Dataset (SAD). (a) Aircraft size distribution. (b) Aircraft aspect ratio distribution.

Figure 9 shows the sample images in SAD. It can be seen that the size of aircraft targets in SAD varies greatly, and the background around some aircraft targets is very complex, which makes the accurate positioning of aircraft difficult. In addition, the background clutter points are easy to confuse with aircraft components, which further aggravates the difficulty of detection.
In this paper, in order to verify our method, we randomly divide the SAD and generate a training set with 2229 images and a test set with 1486 images.

4.2. Evaluation Metrics

For SAR aircraft detection, precision rate and recall rate are usually adopted as evaluation criteria. However, there is usually an opposition between precision and recall, with one increasing and the other decreasing. Therefore, we supplemented F1 score, which is a comprehensive indicator between accuracy and recall. The calculation formulas are as follows:

\[
\text{Precision} = \frac{N_{TP}}{N_{TP} + N_{FP}},
\]

\[
\text{Recall} = \frac{N_{TP}}{N_{TP} + N_{FN}},
\]
\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{9}
\]

where \(N_{TP}\) represents the amount of targets correctly detected, \(N_{FP}\) represents the amount of false alarms, and \(N_{FN}\) indicates the amount of missing aircraft.

### 4.3. Experimental Setup

SFRE-Net uses \(640 \times 640\) image resolution in training and testing and uses random rotation, random flip, and mosaic data-enhancement methods in the training stage. All methods are verified on an NVIDIA RTX3090 GPU. The configuration of experimental parameters is shown in Table 1.

**Table 1.** Hyperparameter settings during model training.

<table>
<thead>
<tr>
<th>Hyper Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.001</td>
</tr>
<tr>
<td>Momentum</td>
<td>0.937</td>
</tr>
<tr>
<td>Decay</td>
<td>0.0005</td>
</tr>
<tr>
<td>Warmup</td>
<td>3</td>
</tr>
<tr>
<td>Batch size</td>
<td>32</td>
</tr>
<tr>
<td>Epochs</td>
<td>200</td>
</tr>
</tbody>
</table>

### 4.4. Comparison of Results with the State of the Art

To thoroughly verify our algorithm, we compared it with the state-of-the-art algorithms, shown in Table 2. Compared with other anchor-based or anchor-free algorithms, SFRE-Net has achieved a top level in multiple evaluation metrics, especially in recall rate. This is because SFRE-Net fully considers the scattering characteristics of aircraft targets in SAR images, uses cascaded transformer structure to capture the correlation between target scattering points, and can reduce the semantic conflict between different-scale feature maps through adaptive fusion method, and the context attention module further improves the positioning accuracy of target points. The detection results of SFRE-Net and other methods are shown in Figure 10. Currently, there are no publicly available SAR aircraft-detection algorithms based on deep learning. To facilitate other scholars’ research, we open-source our SFRE-Net at https://github.com/hust-rslab/SFRE-Net (accessed on 8 March 2022).

**Table 2.** Comparison of detection results between different algorithms on SAD.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>(F_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>faster R-CNN [33]</td>
<td>0.875</td>
<td>0.929</td>
<td>0.901</td>
</tr>
<tr>
<td>Cascade R-CNN [34]</td>
<td>0.926</td>
<td>0.911</td>
<td>0.918</td>
</tr>
<tr>
<td>FCOS [42]</td>
<td>0.865</td>
<td>0.891</td>
<td>0.878</td>
</tr>
<tr>
<td>RetinaNet [36]</td>
<td>0.923</td>
<td>0.881</td>
<td>0.902</td>
</tr>
<tr>
<td>CenterNet [41]</td>
<td>0.934</td>
<td>0.922</td>
<td>0.928</td>
</tr>
<tr>
<td>YOLOX-s [53]</td>
<td>0.928</td>
<td>0.912</td>
<td>0.920</td>
</tr>
<tr>
<td>YOLOV5s [39]</td>
<td>0.940</td>
<td>0.914</td>
<td>0.927</td>
</tr>
<tr>
<td>SFRE-Net (ours)</td>
<td>\textbf{0.944}</td>
<td>\textbf{0.945}</td>
<td>\textbf{0.944}</td>
</tr>
</tbody>
</table>
Figure 10. Detection results of different algorithms. The green boxes, pink boxes, red boxes, and yellow ellipses denote the correct predictions, missing objects, false alarms, and entire aircrafts, respectively.

4.5. Ablation Studies

In this paper, we choose YOLOV5s as the basic framework to compare the effects of different feature fusion pyramids on aircraft detection in SAR images, as shown in Figure 4. To ensure the fairness of comparative experiments, FEM and TRsB modules are not added
to the FAFP structure. These four feature fusion pyramid structures only use the modules of YOLOV5s. In SAR aircraft detection, the effects of different feature fusion pyramids are shown in Table 3.

Table 3. Comparison effects of different feature fusion pyramids, in which the FEM and TRsB modules are not used in FAFP.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPN [29]</td>
<td>0.931</td>
<td>0.916</td>
<td>0.923</td>
</tr>
<tr>
<td>PANet [30]</td>
<td>0.940</td>
<td>0.914</td>
<td>0.927</td>
</tr>
<tr>
<td>BiFPN [45]</td>
<td>0.936</td>
<td>0.920</td>
<td>0.928</td>
</tr>
<tr>
<td>FAFP (Ours)</td>
<td>0.940</td>
<td>0.921</td>
<td>0.930</td>
</tr>
</tbody>
</table>

Table 3 shows that compared with other feature fusion pyramid structures, FAFP has made certain advantages in performance. Therefore, we took FAFP as the feature fusion part of SFRE-Net and performed ablation experiments on other proposed structures; see Table 4.

The feature-enhancement module (FEM) aggregates feature maps of different receptive fields, which enhances the semantic representation of feature fusion and improves recall rate to some extent. The transformer block (TRsB) captures the correlation between discrete points, facilitates the integrity detection of aircraft targets, and further improves the precision rate and recall rate. The context attention-enhancement module (CAEM) captures information from different receptive fields by a dilated convolution pyramid, then enhances the focus of target feature points by the coordinate attention mechanism while suppressing background clutter points. Finally, it aggregates feature maps from different receptive fields as output by an adaptive fusion method, which greatly enhances the positioning ability of aircraft targets in SAR images. Table 4 shows that CAEM helps to improve the precision and recall rate.

Table 4. Comparison results of different feature fusion pyramids, in which the FEM and TRsB modules are not used in FAFP.

<table>
<thead>
<tr>
<th>FEM/FEM-s</th>
<th>TRsB</th>
<th>CAEM</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>0.940</td>
<td>0.921</td>
<td>0.930</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>0.940</td>
<td>0.926</td>
<td>0.933</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>0.941</td>
<td>0.939</td>
<td>0.940</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.944</td>
<td>0.945</td>
<td>0.944</td>
</tr>
</tbody>
</table>

5. Discussion

This section further discusses the proposed method and demonstrates the effectiveness of SFRE-Net from several aspects through quantitative analysis.

(1) Location of TRsB: Due to the complexity of multi-scale feature fusion, the placement of the transformer block (TRsB) is also worth exploring. Table 5 shows the contribution of TRsB at different locations, where the other components of SFRE-Net remain unchanged, only changing the location of TRsB. F3 → FEM means that TRsB is placed between F3 and FEM, as shown in Figure 2. We can see that in the multi-scale feature fusion, the effect of placing TRsB behind the layer with deep depth and large feature map size is better. The feature map output from F3 layer has a large size and richer feature information, which is helpful for TRsB to capture the correlation between feature points.
Table 5. Effects of TRsB at different locations.

<table>
<thead>
<tr>
<th>Position</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>F5 → F4</td>
<td>0.940</td>
<td>0.941</td>
<td>0.940</td>
</tr>
<tr>
<td>F4 → F3</td>
<td>0.942</td>
<td>0.942</td>
<td>0.942</td>
</tr>
<tr>
<td>F3 → FEM</td>
<td>0.944</td>
<td>0.945</td>
<td>0.944</td>
</tr>
</tbody>
</table>

(2) Types of attention modules in CAEM: The attention mechanism can make the model focus on the region of interest in the image. For SAR aircraft detection, using attention mechanisms is helpful to distinguish between target and background. In order to explore the effects of different attention mechanisms in SAR aircraft detection, we compared several typical attention modules in CAEM, as shown in Table 6, where the other components of SFRE-Net remain unchanged, only changing the types of attention modules in CAEM. Squeeze-and-excitation (SE) [54] pays attention to the importance of channels in the feature map, but ignores the information in spatial direction. Bottleneck attention module (BAM) [55] and convolutional block attention module (CBAM) [56] consider both channel and spatial information, but the local convolution is still used for attention calculation in the specific implementation process. Since aircraft targets in SAR images appear as bright discrete points, using the local convolution to calculate the attention easily introduces too much background information, which is not conducive to the accurate positioning of aircraft scattering points. Coordinate attention (CA) [51] decouples the spatial information into two independent feature maps through two 1D global average pooling operations, one with horizontal information and the other with vertical information. The two sub-attention maps obtained have long-distance dependence information in a specific direction. Finally, the two sub-attention maps are multiplied to obtain a global attention map. The coordinate attention mechanism avoids the shortage of local convolution. Each feature point in the output feature map takes advantage of the long distance information in the horizontal and vertical directions, which is helpful to associate with other scattering points and distinguishes between background clutter and aircraft scattering points.

Table 6. Effects of different types of attention modules in CAEM on SAR aircraft detection.

<table>
<thead>
<tr>
<th>Types</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE [54]</td>
<td>0.939</td>
<td>0.94</td>
<td>0.939</td>
</tr>
<tr>
<td>CBM [55]</td>
<td>0.941</td>
<td>0.940</td>
<td>0.940</td>
</tr>
<tr>
<td>CBAM [56]</td>
<td>0.942</td>
<td>0.941</td>
<td>0.941</td>
</tr>
<tr>
<td>CA [51]</td>
<td>0.944</td>
<td>0.945</td>
<td>0.944</td>
</tr>
</tbody>
</table>

(3) Overall effects of SFRE-Net: The detection speed of the object-detection algorithm attracts more and more attention from scholars. In order to comprehensively analyze our method, we made statistics on the average inference time of different algorithms on the test set, as shown in Figure 11. SFRE-Net has the highest detection accuracy, and the detection speed is also at the top level.

To fully demonstrate the effectiveness of SFRE-Net, we visualized the detection confidence maps of SFRE-Net versus baseline YOLOV5s, as shown in Figure 12. Compared with the baseline YOLOV5s, SFRE-Net can locate aircraft targets more accurately and suppress the interference of background clutter points with higher detection confidence.
Figure 11. Average inference time per image for different algorithms.

Figure 12. Detection confidence visualization. (a) Ground Truth of SAD. (b) Detection confidence map for the baseline. (c) Detection confidence map for SFRE-Net. The green boxes and red boxes indicate the correct detections and false alarms, respectively.
6. Conclusions

In this paper, we propose a scattering feature relationship enhancement network (SFRE-Net) to alleviate SAR aircraft-detection problems. Firstly, SFRE-Net proposes a feature-adaptive fusion pyramid (FAFP) structure, which enables the network to independently select useful semantic information through the adaptive weighted addition method to alleviate the semantic conflict in the process of different feature fusion, which is conducive to improve the multi-scale representation ability of the model. In addition, we add the feature-enhancement module (FEM) in FAFP to obtain the feature information of different receptive fields, which enriches the semantic information in feature fusion. Secondly, SFRE-Net adopts a structure of cascaded transformer blocks (TRsB), which can model the correlation of scattering points, and is conducive to the integrity detection of aircraft targets. Finally, to address the problem of complex background interference in SAR images, SFRE-Net designs a context attention-enhancement module (CAEM). CAEM fully considers the discrete characteristics of SAR image targets. Firstly, the dilated convolution pyramid is used to improve the receptive field. Then, the coordinate attention is adopted to enhance the attention of target points and suppress the attention of clutter points, and the feature maps of different information are aggregated and output by adaptive fusion. Many experiments on the Gaofen-3 dataset prove the effectiveness of our algorithm and show that SFRE-Net is superior to the state-of-the-art object-detection algorithms.

Author Contributions: Methodology, P.Z. and T.T.; software, P.Z.; validation, P.Z. and H.X.; formal analysis, P.Z.; investigation, P.Z., H.X., T.T., P.G., and J.T.; writing—original draft preparation, P.Z.; writing—review and editing, T.T. and J.T.; supervision, P.G.; project administration, P.Z. and H.X.; funding acquisition, T.T. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 42071339.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We sincerely thank the editor and reviewers for their constructive comments.

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

References

4. Shu, Y.; Li, W.; Yang, M.; Cheng, P.; Han, S. Patch-based change detection method for SAR images with label updating strategy. Remote Sens. 2021, 13, 1236. [CrossRef]


