Unleashing Channel Potential: Space-Frequency Selection Convolution for SAR Object Detection

Ke Li, Di Wang*, Zhangyuan Hu, Wenxuan Zhu, Shaofeng Li*, Quan Wang
School of Computer Science and Technology, Xidian University, Xi'an, China
{like0413, wxzhu, krudy0323}@stu.xidian.edu.cn, {wangdi, lishaofeng, qwang}@xidian.edu.cn

Abstract

Deep Convolutional Neural Networks (DCNNs) have achieved remarkable performance in synthetic aperture radar (SAR) object detection, but this comes at the cost of tremendous computational resources, partly due to extracting redundant features within a single convolutional layer. Recent works either delve into model compression methods or focus on the carefully-designed lightweight models, both of which result in performance degradation. In this paper, we propose an efficient convolution module for SAR object detection, called SFS-Conv, which increases feature diversity within each convolutional layer through a shunt-perceive-select strategy. Specifically, we shunt input feature maps into space and frequency aspects. The former perceives the context of various objects by dynamically adjusting receptive field, while the latter captures abundant frequency variations and textural features via fractional Gabor transformer. To adaptively fuse features from space and frequency aspects, a parameter-free feature selection module is proposed to ensure that the most representative and distinctive information are preserved. With SFS-Conv, we build a lightweight SAR object detection network, called SFS-CNet. Experimental results show that SFS-CNet outperforms state-of-the-art (SoTA) models on a series of SAR object detection benchmarks, while simultaneously reducing both the model size and computational cost.

1. Introduction

Synthetic aperture radar (SAR) is an active microwave imaging remote sensing device that can capture the Earth's surface around the clock and in all weather conditions. Due to the unique imaging mechanism and rich electromagnetic scattering characteristics, SAR images are widely used in ocean monitoring, resource exploration, land cover classification and disaster investigation [3, 38, 39, 44]. With the development of Deep Convolutional Neural Networks (DCNNs) and the maturity of SAR imaging technology, an increasing number of DCNN-based methods have demonstrated excellent performance in SAR object detection task. However, their success heavily relies on intensive computing and storage resources, which poses serious challenges to their deployment in resource-limited environment.

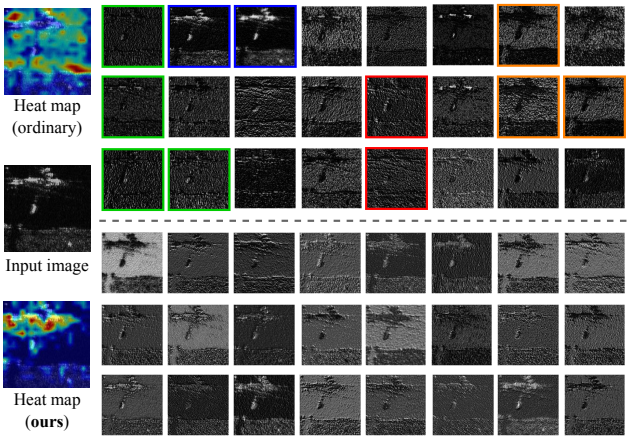

Figure 1. Visualization of feature maps. On the left are the input image and attention heat maps for ordinary convolution and SFS-Conv. Numerous feature maps from ordinary convolutions exhibit much pattern similarities, where four similar feature map group examples are annotated with the same-colored boxes (above). In contrast, the feature maps of SFS-Conv demonstrate greater diversity and distinctiveness (blow).

To address these challenges, researchers have explored various types of model compression strategies and network designs to improve the efficiency of SAR object detection [34, 39, 45]. The former includes network pruning, knowledge distillation, low-rank decomposition, and parameter sparsity. These compression techniques are identified as post-processing methods, thus their performance is typically limited by the quality of the original model. Network design is another approach aimed at reducing the inherent redundancy of dense model parameters and developing a lightweight network. Recent works attempt to use

*Corresponding author

groupwise convolution (GWC) [18], pointwise convolution (PWC) [31] and depthwise convolution (DWC) [33] to build lightweight SAR object detection models [9, 23, 38, 39, 44], so as to enhance detection performance.

Despite the advances, these general convolutions are not initially designed for SAR object detection task. Since SAR images are typically captured at high resolutions from an overhead perspective. Particularly, most objects in SAR images are small, and often obscured by speckle noise. Therefore, it is challenging to identify objects based on appearance alone. Instead, surrounding environment of objects can offer valuable cues for recognition, such as object shape, orientation, and other characteristics [24]. In addition, the imaging principle of SAR relies on recording the interaction between the radar system and objects, which results in the formation of echo signals. Frequency domain analysis can decompose these echo signals into a series of frequency components, where each component represents distinct scattering characteristics exhibited by objects. According to the above points, we conclude two important **priors** for SAR object detection: *i)* Object-adaptive receptive fields facilitate accurate identification. As objects in SAR images exhibit diverse scales, object detectors with fixed receptive fields may yield incorrect classification results. *ii)* Frequency features play a pivotal role in SAR object detection. SAR imaging is often susceptible to intricate background interference, making it difficult to distinguish object features from clutter noise based solely on spatial information.

Inspired by the aforementioned discussions, we design a shunt-perceive-select strategy to fully exploit the potential of channels, with the goal of enhancing the discriminative feature representation ability while reducing the parameter quantity and computational complexity. To this end, we devise a novel **S**pace-**F**requency **S**elective **Conv**olution (SFS-Conv) module tailored for SAR object detection, which consists of three units, a spatial perception unit (SPU), a frequency perception unit (FPU), and a channel selection unit (CSU). The input features are first shunted into space and frequency aspects, and then separately fed into the SPU and FPU to perceive the object's distinctive features such as position, orientations and texture in a more fine-grained manner. Specifically, SPU employs dynamically adjustable convolution kernels to adapt to the abundant context of different objects, thereby effectively modeling relationships between objects and their surrounding environment. The core of FPU is the fractional Gabor transformer, which is used to extract high-frequency texture features at multiple scales and orientations and suppress speckle noise in SAR image. Subsequently, CSU adaptively and selectively fuses the outputs of SPU and FPU to preserve the most representative features in a parameter-free manner. Building upon SFS-Conv, we propose a lightweight SAR object detection network, termed as SFS-CNet, which enhances the representation capacity of the object features.

To the best of our knowledge, this is the first convolutional approach capable of extracting distinctive information in both spatial and frequency dimensions within a single convolutional layer. Most existing SAR object detection methods achieve this by introducing additional spatial or frequency attention modules. However, these methods inevitably increase model complexity, leading to feature redundancy. In contrast to these methods, SFS-Conv enhances the discriminative representation and diversity of convolutional features without increasing the number of channels. In such way, it maintains fewer model parameters and floating point operations (FLOPs). Extensive experiments on benchmark datasets demonstrate that the proposed SFS-Conv reduces redundancy in standard convolutions by emphasizing more representative features. Our contributions are summarized as follows:

- We propose a space-frequency selection convolution, called SFS-Conv, which employs a shunt-perceive-select strategy to enhance the diversity and distinctiveness of features within a convolutional layer while diminishing redundancy in the channel dimension.
- We propose a novel network for SAR object detection based on SFS-Conv, termed as SFS-CNet. It achieves remarkable performance with only 18% parameters and 24% FLOPs compared with the state-of-the-art (SoTA) YOLOv8s.
- Extensive experiments and ablation studies validate the efficacy of our SFS-Conv, achieving accuracies of 96.2% on HRSID [36], 89.7% on SAR-AIRcraft-1.0 [35] and 99.6% on SSDD [20], respectively.

2. Related work

2.1. SAR Object Detection Methods

Two-stage SAR object detectors usually rely on the RCNN [28] framework, which consists of a region proposal network (RPN) and several detection heads. The RPN proposes high-quality regions of interest (RoIs) from the backbone features, while detection heads are responsible for object classification and bounding box regression. The main efforts of two-stage methods are devoted to generating better region proposal, for which several variations on the RCNN framework have been proposed [17, 20, 22]. DAPN [3] uses an attention mechanism to highlight salient features and improve the modeling of multi-scale objects. FFCV [14] introduces fast nonlocal mean (FNLM) filter to reduce background noise and enhance the structural information of the target slice. However, two-stage approaches require filtering a large number of candidate boxes, resulting in large time and computing overhead.

One-stage detection framework does not rely on the pro-

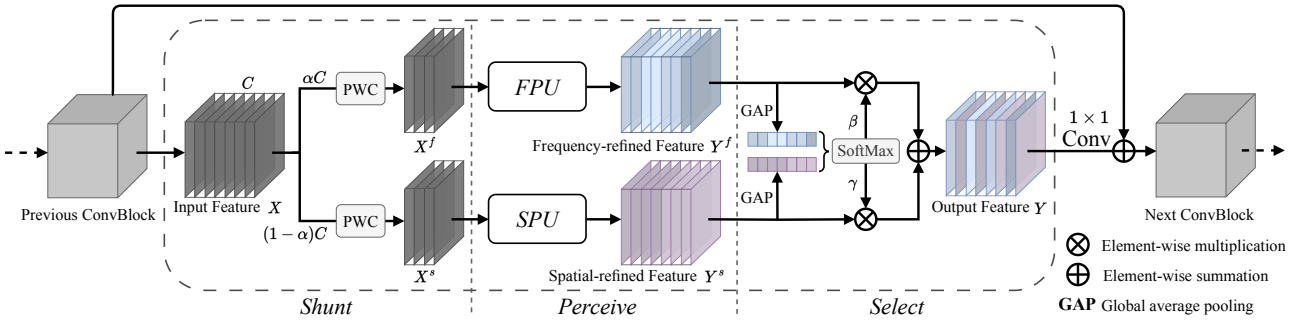

Figure 2. The proposed SFS-Conv module.

posed anchor points or regions but classifies and regresses oriented bounding boxes directly from densely sampled grid anchors. LMSD-YOLO [9] leverages DSC [12] to adaptively fuse the backbone's extracted multi-scale features for more accurate predictions. To effectively mitigate sea clutter in SAR images, YOLO-FA [42] introduces a frequency attention module to generate adaptive frequency weights before spatial feature extraction. CoLD [34] employs optical object detection network as a teacher to guide SAR object detection network, improving the localization quality of oriented bounding boxes.

Although the aforementioned methods have shown promising results in SAR object detection tasks, they either introduce numerous multi-scale and attention-related modules to ensure detection accuracy or blindly pursue lightweight and detection speed, which may lead to a decline in detection performance. In contrast, we focus on redesigning the convolution to extract diverse and representative information in both spatial and frequency domains, rather than simply stacking modules. This approach allows us to achieve optimal performance while maintaining extremely low inference times.

2.2. Efficient Convolution Design

Convolution design aims to reduce the inherent redundancy of dense parameters in deep models and further develop lightweight network architectures. AlexNet [19] introduces the concept of GWC, which divides channels into different groups for parallel convolution operations to improve computational efficiency. Inception [31] adopts split-transform-merge strategy and achieves low theoretical complexity with compelling accuracy. Meanwhile, some novel convolutional filters like PWC [31] and DWC [33] are widely used for efficient model design by reducing inter-channel density connections. Evolved from Inception, ResNeXt [37] replaces traditional convolutions with sparsely connected group convolutions to reduce redundancy in inter-channel connectivity. To further reduce connection density, MobileNet [12] takes advantage of depthwise separable convolution (DSC), which utilizes DWC for spatial information extraction and PWC for channel information fusion successively. Moreover, ShuffleNet [46] adopts GWC on 1x1 convolutions followed by channel shuffle operation.

However, as shown in Fig. 1, numerous feature maps exhibit pattern similarity, indicating that standard convolution operations may produce redundant information. GhostNet [10] and SPConv [43] also point out redundancy among feature maps, both employing lightweight operations to learn channel-wise redundancy. SCConv [21] proposes a two-step compression approach to jointly reduce spatial and channel redundancy in convolutional layers. The above methods primarily suppress redundant information in feature maps and ignore to generate more discriminative and diverse features. Additionally, SAR images usually contain a significant amount of speckle noise, and it is insufficient to focus only in the spatial or channel dimension. Our SFS-Conv extracts diverse features in spatial and frequency dimensions and further selects more representative features, resulting in improved performance.

3. Method

3.1. Ordinary Convolution

Let $X \in \mathbb{R}^{L\times H\times W}$, $Y \in \mathbb{R}^{M\times H'\times W'}$ denote the input and convolved output tensors with L input channels and M output channels respectively. Eq. (1) gives the $k \times k$ convolutional kernels operation [1] of value $Y_{M,h',k'}$ at the spatial position (h', w'):

$$Y_{M,h',w'} = \sum_{z=0}^{L-1}\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} X_{z,h+i,w+j} * K_{z,i,j,M}, \qquad (1)$$

where $X_{z,h,w}$ represents the value in X at spatial position (h, w) and channel z, $*$ is an element-by-element product operation, and $K \in \mathbb{R}^{L\times k\times k\times M}$ represents square $k \times k$ kernels, which are used to convolve L input channels into M output channels for feature extraction. The convolution operation is both local and linear. As each convolutional

[1] To simplify the notation, we omit the bias term.

kernel slides across the entire input, it operates in a local manner without considering global context or prior knowledge. This can result in the repetitive extraction of unnecessary or similar features in feature maps.

3.2. Space-Frequency Selective Convolution

The proposed SFS-Conv module is illustrated in Fig. 2, which consists of three units, a spatial perception unit (SPU), a frequency perception unit (FPU), and a channel selection unit (CSU). Specifically, through SPU and FPU, the SFS-Conv obtains spatial-refined features Y^s and frequency-refined features Y^f, respectively. Then, representative features Y are reserved through CSU operation.

Before feeding the input feature maps $X \in \mathbb{R}^{C\times H\times W}$ into SPU and FPU, we first partition X into two aspects in a ratio of α. One is used to represent the spatial aspects, denoted as $X^s \in \mathbb{R}^{(1-\alpha)C\times H\times W}$, providing spatial information. The other is employed for the frequency aspects, denoted as $X^f \in \mathbb{R}^{\alpha C\times H\times W}$, to complement frequency characteristics. Subsequently, we use two 1×1 PWCs to individually adjust X^s and X^f, making them more suitable for the next step of feature extraction in spatial and frequency dimensions.

3.2.1 SPU for Spatial Context Representation

Referring to the aforementioned *priors (i)*, as described in Sec. 1, we introduce a spatial perception unit (SPU) that dynamically models contextual information across various scales. Specifically, we first partition the feature map channels, and then adopt multiple kernels of different sizes to obtain multi-scale features. In addition, we construct hierarchical residual connections between kernels to further expand the receptive field for each convolution layer. As shown in Fig. 3, we evenly split the spatial part X^s into n feature map groups X^s_g, where $g \in \{1, 2, \ldots, n\}$. Each group X_g has the same spatial dimension as X^s but has only $1/n$ channel length. For each X^s_g, there exists a corresponding convolution kernel with size of k_g, denoted as $K_g \in \mathbb{R}^{C\times k_g\times k_g\times C}$. For simplicity, we omit the summation operations of each dimension in convolution. Therefore, the output Y^s_g of the K_g can be represented as:

$$Y^s_g = \begin{cases} X^s_g * K_g, & g = 1 \\ \left(X^s_g + Y^s_{g-1}\right) * K_g, & 1 < g \leq n \end{cases}. \tag{2}$$

Specifically, for the $g-$th group of convolution kernels, the expansion of kernel size k_g and the receptive field RF_g are defined as follows:

$$\begin{aligned} k_{g+1} &= k_g + 2,\ k_1 = 3, \\ RF_{g+1} &= RF_g + (k_{g+1} - 1)\,. \end{aligned} \tag{3}$$

As each convolution K_g could receive feature information from all feature groups $\{X^s_i, i < g\}$, the output Y^s_g exhibits a larger receptive field than preceding group $\{Y^s_i, i < g\}$. This multi-scale manner is conducive to capturing rich and valuable visual context, facilitating accurate identification. To better fuse information at different scales, we concatenate all groups along the channel and pass them through a 1×1 convolution, which is denoted as:

$$Y^s = \sum_{i=0}^{C-1} \left[Y^s_1, Y^s_2, \ldots, Y^s_g\right]_i * K_{i,C}, \tag{4}$$

where $[\cdots]$ represents concatenate operation, Y^s is the spatial-refined features. In SPU, we use g as a control parameter of the scale dimension. A larger g can allow channels within the same convolutional layer to respectively account for multi-granularity features, thereby reducing similar features and enhancing feature representation capability.

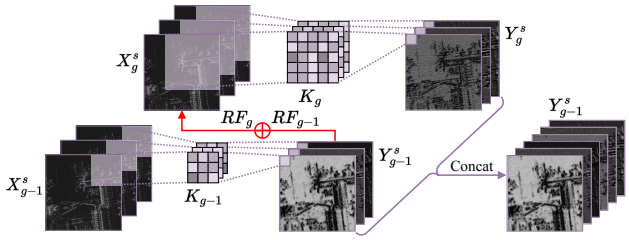

Figure 3. The proposed Spatial Perception Unit (SPU).

3.2.2 FPU for Frequency Analysis

Referring to the aforementioned *priors (ii)*, as described in Sec. 1, we introduce a Frequency Perception Unit (FPU). In SAR scenes, objects typically exhibit characteristics across multiple scales and orientations, making traditional convolution kernels potentially inflexible in feature extraction and leading to feature redundancy. For example, in a scenario where an object undergoes rotation, features extracted by traditional convolution kernels may struggle to adapt to the new direction, necessitating more kernels to cover features in different directions. The Gabor Transform (GT) has demonstrated superiority in object recognition, particularly in cases involving frequent changes in scale and rotation [26]. Additionally, the Fractional Fourier Transform (FrFT) has proven effective in mitigating Doppler Shifts caused by multiple moving objects [30]. Based on this, we attempt to incorporate the Fractional Gabor Transform (FrGT) [47] into SAR object detection. This transformation guides convolution kernels in extracting high-frequency texture features at multiple scales and orientations, effectively suppressing speckle noise in SAR images.

Fractional Gabor Transformer. FrGT replaces the conventional Fourier transform with the FrFT in the definition of the GT [47]. For signal $s(x)$, Eq. (5) provides the

definition of standard FrGT (one-dimensional and continuous), whose mathematical definition is given by:

$$\begin{aligned} G_s^{\alpha}(p,q) &= \int s(x)\bar{g}(x-q)B(p,x,\alpha)\mathrm{d}x, \\ B(x_1,x_2,\alpha) &= \sqrt{1-i\cot\alpha}\exp\{i\pi[(x_1^2+x_2^2)\cot\alpha \\ &\quad -2x_1x_2\csc\alpha]\}, \end{aligned} \tag{5}$$

where $g(\cdot)$ is a window function, $\bar{}$ stands for the complex conjugate, $B(x_1,x_2,\alpha)$ is a transform kernel, $\alpha = P\pi/2$ is the transform angle, P is the transform order, and p and q are the coordinates in the space and the FrFT domain, respectively. In order to modulate Conv2D, we extend standard FrGT to 2-D discrete space [26, 41]. For an image $f(x,y) \in \mathbb{R}^{C\times H\times W}$, the FrGT can be formulated as:

$$\begin{aligned} G_f^{\alpha}(x,y,u,v) = \sum_{i=0}^{H-1}\sum_{m=0}^{U-1} B(i,\frac{m}{UT_1},\alpha)\bar{g}(i-m) \\ \left[\sum_{j=0}^{W-1}\sum_{n=0}^{V-1} f(i,j)B(j,\frac{n}{VT_2},\alpha)\bar{g}(j-n)\right], \end{aligned} \tag{6}$$

where x, y and u, v are the coordinates in the space and the FrFT domain, H and W are the size of the input image, U and V are the numbers of samples in the fractional domain, T_1 and T_2 are sampling intervals.

Convolutional Fractional Gabor Kernels (FrGK). To empower standard convolutional kernels within the same convolutional layer to capture texture features with diverse scales and orientations, we employ FrGT filters to modulate ordinary convolutional kernels, as described by [26]:

$$K_{i,u}^{v} = K_{i,o} * G(u,v), \tag{7}$$

where $K_{i,o}$ is a learned $k \times k$ kernel with i input channels and o output channels, $G_{u,v}$ represents a group of FrGT filters with different orientations and scales. Due to the rotation equivariance property of the FrGT, we can share parameters across different orientations of FrGK $K_{i,u}^{v}$ in v-scale, where $u \in \{0, U-1\}$ and $v \in \{0, V-1\}$. To facilitate implementation, we set the number of FrGK channels to N times the number of Gabor orientations U. Thus, a learned kernel $K_{i,U\cdot N}^{v}$ with i input channels and $U \cdot N$ output channels is obtained by concatenating N FrGK kernels with the same scale but different orientations, which is denoted as:

$$K_{i,U\cdot N}^{v} = \left[K_{i,U_0}^{v}, K_{i,U_1}^{v}, \cdots, K_{i,U_{N-1}}^{v}\right]. \tag{8}$$

For input features X^f with C channels, we divide them into V groups X^{f_v}, $v \in \{0, V-1\}$. Each X^{f_v} requires $N = C/VU$ convolutional kernels to generate the corresponding frequency features Y^{f_v}, as indicated below:

$$Y^{f_v} = X^{f_v} * K_{i,C/V}^{v}, \tag{9}$$

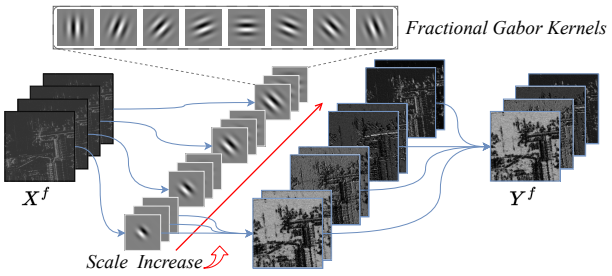

Figure 4. The proposed Frequency Perception Unit (FPU).

Finally, we concatenate all Y^{f_v} to obtain the frequency-refined features $Y^f = [Y^{f_0}, Y^{f_1}, \cdots, Y^{f_{V-1}}]$.

3.2.3 CSU for Parameter-Free Fusion

So far, we have obtained the spatial-refined features Y^s and the frequency-refined features Y^f. Since these two kind of features originate from different input channels, a fusion method is needed to adaptively select more discriminative features. Specifically, we first use global average pooling (GAP) to collect globally spatial and frequency information with channel-wise statistics $S^n \in \mathbb{R}^{C\times 1\times 1}, n \in \{s, f\}$, which can be represented as:

$$S^n = GAP(Y^n) = \frac{1}{H\times W}\sum_{i=0}^{H-1}\sum_{i=0}^{W-1} Y_{i,j}^{n}. \tag{10}$$

Next, we stack S^s and S^f together and use channel-wise soft attention operation to generate feature selectivity weights $\gamma, \beta \in \mathbb{R}^C$, which are defined as follows:

$$\gamma = \frac{e^{S^s}}{e^{S^s}+e^{S^f}},\ \beta = \frac{e^{S^f}}{e^{S^s}+e^{S^f}}. \tag{11}$$

Finally, under the guidance of feature selectivity weights γ and β, the final output Y can be obtained by merging Y^s and Y^f as follows:

$$Y = \gamma Y^s + \beta Y^f. \tag{12}$$

Overall, by arranging SPU, FPU and CSU through the shunt-perceive-select strategy, the proposed SFS-Conv is established, which can extract informative and discriminative object features in SAR images.

3.3. SFS-CNet Architecture

With SFS-Conv, we present a lightweight SAR object detector named SFS-CNet, and its framework is depicted in Fig. 5. Specifically, we construct the backbone network by continuously stacking four CBR and SFS-Conv modules, where CBR is a downsampling module composed of a 3×3 convolution with a batch normalization (BN) layer

Methods	Params. (M)	FLOPs (G)	Infer. Time (ms/img)	All					Offshore					Inshore				
				R	P	AP_{50}	AP_{75}	$F1$	R	P	AP_{50}	AP_{75}	$F1$	R	P	AP_{50}	AP_{75}	$F1$
Faster R-CNN [28]	41.30	138.5	18.0	81.5	86.1	86.5	73.6	83.7	95.4	96.8	97.7	94.4	96.1	71.1	77.4	78.4	58.4	74.1
Mask R-CNN [11]	43.99	188.4	24.9	82.3	87.1	87.9	75.0	84.6	96.2	97.3	98.3	94.1	96.7	71.7	78.2	79.2	60.8	74.8
Cascade R-CNN [1]	69.15	159.2	23.1	81.8	86.2	86.6	76.8	83.9	95.6	96.8	98.1	95.3	96.2	72.4	78.8	79.5	69.3	75.5
Cascade Mask R-CNN [1]	77.05	-	-	82.6	88.0	88.3	77.2	85.2	96.3	97.4	98.5	95.3	96.8	73.2	79.3	80.0	64.7	76.1
HTC [2]	80.02	-	-	82.4	86.9	87.6	78.5	84.6	96.9	97.4	98.7	**96.1**	97.1	74.8	81.0	82.1	**71.2**	77.8
HRSDNet [36]	91.03	-	-	-	-	89.3	79.8	-	-	-	98.6	96.0	-	-	-	81.3	68.3	-
RetinaNet [25]	36.25	139.2	16.4	77.2	82.9	83.7	66.5	79.9	95.6	96.2	97.6	92.5	95.9	61.5	69.6	69.3	42.7	65.3
FCOS [32]	32.11	123.3	17.4	76.8	80.1	80.7	57.3	78.4	95.8	96.7	97.8	87.5	96.2	57.3	64.3	64.5	33.7	60.6
CenterNet [5]	32.13	123.0	18.4	82.0	86.4	87.0	64.9	84.1	95.9	96.7	97.9	90.6	96.3	71.3	76.7	77.6	46.0	73.9
YOLO-FA [42]	6.86	-	14.7	87.6	93.1	93.5	-	90.3	-	-	-	-	-	-	-	-	-	-
YOLOv3 [27]	61.50	77.4	13.2	84.6	91.0	91.9	74.0	87.7	96.1	97.3	98.2	75.6	96.7	79.0	68.5	75.7	21.0	73.4
YOLOX-Nano [7]	0.90	0.5	15.9	72.8	79.4	80.1	54.8	76.0	94.3	95.7	96.9	87.4	95.0	72.5	79.9	80.1	26.7	76.0
YOLOv5n [15]	1.92	4.5	6.3	84.3	90.7	91.4	70.4	87.4	96.0	97.5	98.2	80.5	96.7	77.9	68.6	75.7	47.1	73.9
YOLOv5s [15]	7.21	16.5	6.4	89.3	94.2	95.4	83.3	91.7	**97.2**	98.1	98.9	93.3	**97.6**	**79.9**	85.5	86.9	65.7	82.6
YOLOv8n [16]	3.01	8.9	12.2	86.9	93.0	93.7	80.2	89.8	96.2	97.8	98.8	94.2	97.0	72.6	83.6	80.3	57.1	77.7
YOLOv8s [16]	10.65	28.4	14.1	90.8	95.0	**96.2**	**87.2**	92.9	96.6	**98.6**	**99.2**	96.0	**97.6**	79.1	88.1	87.3	70.2	**83.3**
ROI Transformer [4]	55.03	-	50.5	84.0	-	79.7	49.4	-	94.7	97.4	90.7	-	96.0	56.1	80.1	58.0	-	66.0
PVT-SAR [48]	31.43	-	-	-	-	-	-	-	96.6	98.0	90.8	-	97.3	73.2	74.9	72.0	-	74.0
SFS-CNet **(ours)**	1.86	6.9	8.6	88.8	95.1	95.7	84.5	91.8	96.5	98.1	98.8	94.9	97.3	78.0	84.9	85.9	63.7	81.3
SFS-CNet † **(ours)**	1.86	6.9	8.6	**90.9**	**95.1**	**96.2**	86.8	**93.0**	96.4	98.5	99.1	95.5	97.4	78.3	**88.9**	**87.5**	69.5	**83.3**

Table 1. Comparison of SFS-CNet and SoTA methods on **HRSID** data set. The best and second best performance are highlighted in **bold** and underline. '†' represents using the OGL strategy.

and a rectified linear unit (ReLU). Our detector takes a $640 \times 640 \times 3$ image as input, and after first downsampling, the number of channels becomes $C = 32$. For the CBR module, the space reduction ratios of the four stages are all 2. Then, we perform two upsampling operations (*i.e.,* CBR, Upsample, Concat) to achieve detection at a larger resolution of $\frac{H}{8} \times \frac{W}{8} \times 4C$, where H and W represent the height and width of the input image.

Object-level Gradient-induced Learning In SAR imagery, the gradient of the image reflects directional changes in scatter intensity between neighboring positions, providing crucial information about object features and structures. Inspired by the success of DGNet [13] in camouflaged object detection, we propose an object-level gradient-induced learning (OGL) strategy to emphasize detailed object textures. Instead of directly using the raw gradient map, we apply the Canny edge detector within the bounding box area where the objects are located in SAR image, formulated as: $g_o = \mathrm{Canny}(\mathrm{bbox}(X))$. Unlike previous methods, OGL not only preserves gradient cues at object boundaries and interior regions but also considers the retention of surrounding gradients near the object. This refinement allows SFS-CNet to better understand contextual information, leading to improved accuracy and robustness, particularly in dense or complex backgrounds. In addition, OGL is employed exclusively during training, thereby avoiding an increase in computational complexity and time during inference.

4. Experiments

4.1. Implementation Details

In our experiment, we present the results of object detection models on HRSID, SAR-Aircraft-1.0 and SSDD data sets. To assess the effectiveness of our model, we utilize four

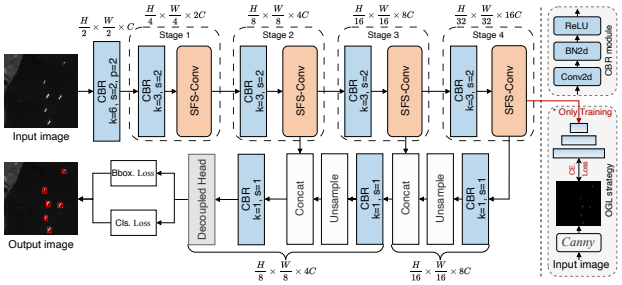

Figure 5. The proposed SFS-CNet framework.

metrics commonly employed in prior research: Recall (R), Precision (P), F1-score ($F1$), and Average Precision (AP). Additionally, we provide details on parameters, FLOPs and the inference time (Infer. Time) for a single image. All reported FLOPs and inference times are calculated using a 640×640 image input with batch size $= 1$. We use Xavier init [8] to randomly initialize parameters of other components in our network. The proposed SFS-CNet is optimized for 300 epochs using the SGD [29], with an initial learning rate of 0.01, and weight decay and momentum set to 5×10^{-4} and 0.9, respectively. Training is conducted on a single RTX3090 GPU.

4.2. Main Results

Results on HRSID. We evaluated the performance of our SFS-CNet against 18 SoTA methods on HRSID data set. The results presented in Tab. 1 demonstrate that our proposed method achieves optimal performance across all metrics when compared to models with similar parameters and FLOPs, such as YOLOv5n and YOLOv8n. In contrast to the larger-sized YOLOv8s, SFS-CNet achieves detection

Methods	mAP	A330		A320/321		A220		ARJ21		Boeing737		Boeing787		other	
		AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}
Faster R-CNN [28]	76.1	85.0	85.0	97.2	87.7	78.5	58.7	74.0	55.2	55.1	42.8	72.9	60.5	70.1	45.4
Cascade R-CNN [1]	75.7	87.4	87.4	97.5	73.9	74.0	49.1	78.0	59.0	54.5	39.1	68.3	57.6	69.1	46.1
RepPoints [40]	72.6	89.8	66.4	97.9	84.9	71.4	49.4	73.0	50.9	55.7	36.6	51.8	41.8	68.4	43.1
SkG-Net [6]	70.7	79.3	66.4	78.2	49.6	66.4	29.8	65.0	37.7	65.1	48.7	69.6	51.6	71.4	41.4
SA-Net [35]	77.7	88.6	88.6	94.3	86.6	90.3	55.0	78.6	59.7	59.7	41.8	70.8	60.4	71.3	47.7
RetinaNet [25]	72.3	92.0	70.1	92.6	58.4	73.0	41.7	63.2	47.1	47.8	25.3	65.4	50.0	67.0	42.3
FCOS [32]	55.2	30.8	29.5	65.6	64.5	60.2	33.0	57.6	35.5	41.9	20.2	46.8	34.3	62.6	33.0
CenterNet [5]	71.1	91.4	69.3	92.3	64.4	70.5	44.0	64.6	45.6	47.3	26.4	65.9	49.7	66.1	41.0
YOLOv3 [27]	83.9	91.8	90.8	96.9	**97.0**	86.5	**69.6**	77.5	61.4	77.0	52.6	76.4	65.8	82.4	57.2
YOLOX-Nano [7]	81.3	95.6	74.7	96.9	74.8	79.7	45.3	78.7	39.5	66.6	39.7	78.2	51.1	73.8	37.7
YOLOv5n [15]	88.2	88.2	83.3	98.9	68.2	84.6	52.5	86.6	56.1	75.0	69.3	95.2	77.6	84.7	54.4
YOLOv5s [15]	89.0	92.1	92.1	98.9	73.1	87.4	60.7	86.4	56.9	76.3	70.2	**96.2**	**86.7**	85.1	59.0
YOLOv8n [16]	88.4	93.1	92.0	97.2	73.1	85.6	56.3	86.1	**66.1**	74.7	70.5	91.1	82.6	83.1	58.1
YOLOv8s [16]	89.6	95.0	95.2	97.7	88.5	**95.8**	60.2	86.6	65.0	**78.9**	**74.2**	90.9	81.8	84.4	59.6
SFS-CNet **(ours)**	88.7	91.4	86.2	97.6	73.9	87.6	58.8	**87.7**	60.9	77.8	71.6	92.4	83.6	**86.6**	**60.8**
SFS-CNet † **(ours)**	**89.7**	**95.9**	**95.9**	**99.3**	74.0	87.9	59.8	86.7	61.3	77.9	69.3	92.9	86.3	85.6	59.6

Table 2. Comparison of SFS-CNet and SoTA methods on **SAR-AIRcraft-1.0** data set. The best and second best performance are highlighted in **bold** and underline. '†' represents using the OGL strategy.

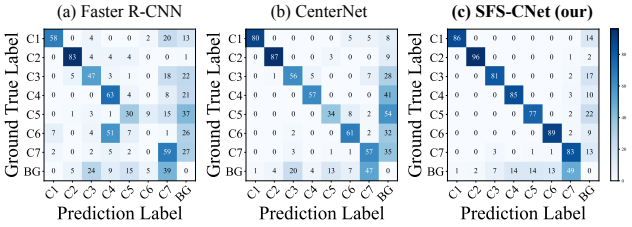

Figure 6. Confusion matrices of different methods. C1-C7 and BG represent A330, A320/321, A220, ARJ21, Boeing737, Boeing787, other and background, respectively.

performance equivalent to or even surpassing it with only **18%** of model size and **25%** of FLOPs, especially in more complex inshore ship detection scenarios. Notably, our SFS-CNet can infer an image in just **8.6** ms, saving **39%** inference time compared to YOLOv8s.

Results on SAR-AIRcraft-1.0. We compare SFS-CNet with 14 SoTAs on SAR-AIRcraft-1.0 data set, as reported in Tab. 2. SFS-CNet† achieves the SoTA with an mAP of **89.7%**, surpassing YOLOv5n and YOLOv8n by **+1.5%** and **+1.3%** mAP, respectively. Additionally, SFS-CNet also demonstrates formidable capabilities in fine-grained recognition. Fig. 6 shows the confusion matrices of the fine-grained recognition results among SFS-CNet with the two-stage method Faster R-CNN [28] and the one-stage method CenterNet [5]. Our confusion matrix exhibits values concentrated along the diagonal, indicating a lower error classification rate.

Results on SSDD. We compare our SFS-Conv against 13 other models on the SSDD dataset, as shown in Tab. 3. The results reveal that our SFS-CNet and SFS-CNet† perform exceptionally well, achieving SoTA AP_{50} scores of **99.4%** and **99.6%** respectively, outperforming all other methods.

Methods	R	P	AP_{50}	$F1$
Faster R-CNN [28]	90.4	87.1	89.7	88.7
Cascade R-CNN [1]	90.8	84.1	90.5	92.4
HRSDNet [36]	91.0	96.5	90.8	93.7
RetinaNet [25]	93.1	94.0	95.0	93.6
FCOS [32]	93.9	94.6	95.3	94.2
CenterNet [5]	93.6	94.3	95.1	93.9
YOLO-FA [42]	95.0	95.2	96.8	95.1
YOLOv3 [27]	94.8	95.5	96.2	95.2
YOLOX-Nano [7]	93.8	95.0	96.2	94.4
YOLOv5n [15]	95.8	97.0	98.0	96.4
YOLOv5s [15]	96.4	97.2	98.3	97.0
YOLOv8n [16]	96.5	97.3	98.1	96.9
YOLOv8s [16]	97.2	98.0	99.4	97.6
SFS-CNet **(ours)**	97.3	**98.2**	99.4	97.7
SFS-CNet † **(ours)**	**97.4**	**98.2**	**99.6**	**97.8**

Table 3. Comparison of SFS-CNet and SoTA methods on **SSDD** data set. The best and second best performance are highlighted in **bold** and underline. '†' represents using the OGL strategy.

4.3. Ablation Study

In this section, we report ablation study results on the HRSID test set to investigate its effectiveness without OGL strategy.

Effect of *Shunt* strategy' component. We conducted experiments to determine the optimal strategy for *Shunt* approach, as described in Tab. 4. The results indicate that emphasizing only spatial or frequency information leads to performance decreases of **0.91%** and **1.18%**, respectively. Setting α to 1/4 achieves the best AP_{50}. When $\alpha = 1/2$, although the accuracy decreases by **0.02%**, the model size decreases by **0.11M**, providing a better trade-off between performance and efficiency. Further increasing the shunt ratio α to 3/4 and 1 also incurs a **0.35%** performance drop.

Effect of *Perceive* strategy' component. To validate the

Shunt	Infer. Time	Params.	AP_{50}
$\alpha = 0$	8.3 ms	1.87 M	94.80 %
$\alpha = 1/4$	8.6 ms	1.97 M	**95.73** %
$\alpha = 1/2$	8.6 ms	1.86 M	95.71 %
$\alpha = 3/4$	9.4 ms	1.73 M	95.38 %
$\alpha = 1$	8.8 ms	1.30 M	94.55 %

Table 4. Effectiveness of the shunt strategy of SFS-Conv.

Perceive		Infer. Time	Params.	AP_{50}
SPU	FPU			
-	-	8.8 ms	2.18 M	90.39 %
✓		8.4 ms	2.17 M	94.66 %
	✓	9.9 ms	2.05 M	94.45 %
✓	✓	8.6 ms	1.86 M	95.71 %

Table 5. Effectiveness of the perceive strategy of SFS-Conv.

Select		Infer. Time	Params.	AP_{50}
CS	SS			
-	-	8.6 ms	1.86 M	94.68 %
	✓	8.8 ms	2.01 M	95.82 %
✓		8.6 ms	1.86 M	95.71 %

Table 6. Effectiveness of the selection strategy of SFS-Conv. CS and SS represent channel selection and spatial selection, respectively.

effectiveness of proposed *Perceive* strategy, *i.e.*, spatial perception unit (SPU), and frequency perception unit (FPU), we conduct ablation experiments, which are summarized in Tab. 5. In the absence of SPU and FPU, visual features are extracted using ordinary 3×3 convolution, resulting in a **5.32%** performance decrease. When only using SPU or FPU, it can be observed that the model already achieves **94.66%** and **94.45%**, respectively, surpassing the **94.2%** reported by YOLOv5s [15]. When combined with SPU and FPU, further improvements can be achieved with **+1.05%** and **+1.26%**, respectively.

Effect of *Select* strategy' component. We also validate the impact of different fusion methods on *Select* strategy. As shown in Tab. 6, a straightforward addition of the two aspects' features yields only **94.68%** AP_{50}. Employing spatial selection (similar in LSKNet [24]) yielded the highest detection accuracy, However, this comes at the cost of increased inference time and parameters by **0.2** ms and **0.15** M, respectively. In contrast, our CSU adopts a parameter-free fusion approach, adaptively selecting more discriminative features from both spatial and frequency aspects, leading to faster inference speed and a lower model size.

4.4. Visualize Results

The visualization results for HRSID and SAR-Aircraft-1.0 data sets are shown in Figure 7. It can be observed that, even in complex scenes, our SFS-CNet successfully detects and recognizes objects, with the network's attention effectively focused on the objects themselves. Notably, SFS-CNet can attend to intricate texture details and frequency variations of objects. Even in scenarios where spatial information is limited, the proposed model maintains a high attention to objects.

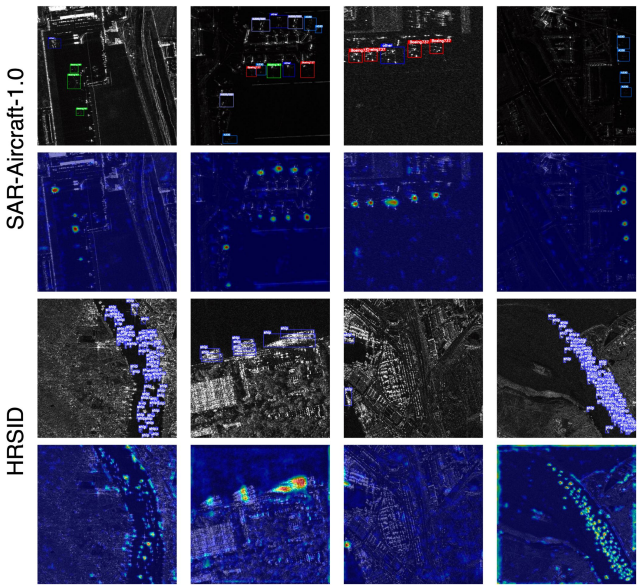

Figure 7. Visualization results on the HRSID and SAR-Aircraft-1.0 data sets.

5. Conclusions

In this paper, we propose the **S**pace-**F**requency **S**elective **Conv**olution (SFS-Conv) for SAR object detection. SFS-Conv employs a shunt-perceive-select strategy to enhance the diversity and distinctiveness of features within a convolutional layer. Through this strategy, we extract informative features from both spatial and frequency aspects, achieving efficient fusion in a parameter-free manner. With SFS-Conv, we propose a lightweight SAR object detection network, termed SFS-CNet. Extensive experiments demonstrate that our proposed lightweight model achieves state-of-the-art performance on competitive SAR benchmarks.

6. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62072354, 62172222 and 62072355, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23084, in part by the Shaanxi Key Research and Development Program under Grant 2024JC-YBQN-0732, in part by the Science and Technology Program of Guangzhou under Grant SL2022A04J00303, and in part by the Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 6, 7

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019. 6

[3] Zongyong Cui, Qi Li, Zongjie Cao, and Nengyuan Liu. Dense attention pyramid networks for multi-scale ship detection in sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):8983–8997, 2019. 1, 2

[4] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 6

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 6, 7

[6] Kun Fu, Jiamei Fu, Zhirui Wang, and Xian Sun. Scattering-keypoint-guided network for oriented ship detection in high-resolution and large-scale sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:11162–11178, 2021. 7

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6, 7

[8] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 6

[9] Yue Guo, Shiqi Chen, Ronghui Zhan, Wei Wang, and Jun Zhang. Lmsd-yolo: A lightweight yolo algorithm for multi-scale sar ship detection. *Remote Sensing*, 14(19):4801, 2022. 2, 3

[10] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020. 3

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[13] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, 2023. 6

[14] Mingda Jiang, Lingjia Gu, Xiaofeng Li, Fang Gao, and Tao Jiang. Ship contour extraction from sar images based on faster r-cnn and chan–vese model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 2

[15] Glenn Jocher. YOLOv5 by Ultralytics, 2020. 6, 7, 8

[16] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 6, 7

[17] Miao Kang, Xiangguang Leng, Zhao Lin, and Kefeng Ji. A modified faster r-cnn based on cfar algorithm for sar ship detection. In *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)*, pages 1–4. IEEE, 2017. 2

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3

[20] Jianwei Li, Changwen Qu, and Jiaqi Shao. Ship detection in sar images based on an improved faster r-cnn. In *2017 IEEE BIGSARDATA*, 2017. 2

[21] Jiafeng Li, Ying Wen, and Lianghua He. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6153–6162, 2023. 3

[22] Yiding Li, Shunsheng Zhang, and Wen-Qin Wang. A lightweight faster r-cnn for ship detection in sar images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2020. 2

[23] Yuankui Li, Xiaoqi Lv, Pingping Huang, Wei Xu, Weixian Tan, and Yifan Dong. Sar ship target detection based on improved yolov5s. In *2021 International Conference on Control, Automation and Information Sciences (ICCAIS)*, pages 354–358, 2021. 2

[24] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection. *arXiv preprint arXiv:2303.09030*, 2023. 2, 8

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 6, 7

[26] Shangzhen Luan, Chen Chen, Baochang Zhang, Jungong Han, and Jianzhuang Liu. Gabor convolutional networks. *IEEE Transactions on Image Processing*, 27(9):4357–4366, 2018. 4, 5

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 6, 7

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 6, 7

[29] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 6

[30] Hong-Bo Sun, Guo-Sui Liu, Hong Gu, and Wei-Min Su. Application of the fractional fourier transform to moving target detection in airborne sar. *IEEE Transactions on Aerospace and Electronic Systems*, 38(4):1416–1424, 2002. 4

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 2, 3

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 6, 7

[33] Vincent Vanhoucke. Learning visual representations at scale. *ICLR invited talk*, 1(2), 2014. 2, 3

[34] Chao Wang, Rui Ruan, Zhicheng Zhao, Chenglong Li, and Jin Tang. Category-oriented localization distillation for sar object detection and a unified benchmark. *IEEE TGARS*, 2023. 1, 3

[35] Zhirui Wang, Yuzhuo Kang, Zeng Xuan, Wang Yuelei, ZangTing, and Sun Xian. Sar-aircraft-1.0: High-resolution sar aircraft detection and recognition dataset. *Journal of Radars*, 12:1–17, 2023. 2, 7

[36] Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access*, 8: 120234–120254, 2020. 2, 6, 7

[37] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

[38] Xiaowo Xu, Xiaoling Zhang, Tianjiao Zeng, Jun Shi, Zikang Shao, and Tianwen Zhang. Group-wise feature fusion r-cnn for dual-polarization sar ship detection. In *2023 IEEE Radar Conference*, pages 1–5, 2023. 1, 2

[39] Xiaowo Xu, Xiaoling Zhang, Tianwen Zhang, and Tianjiao Zeng. Group-wise shuffle attention r-cnn for ship detection in dual-polarization sar images. In *IGARSS 2023*, pages 6410–6413, 2023. 1, 2

[40] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9657–9666, 2019. 7

[41] Baochang Zhang, Yongsheng Gao, Sanqiang Zhao, and Jianzhuang Liu. Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE transactions on image processing*, 19(2): 533–544, 2009. 5

[42] Linping Zhang, Yu Liu, Wenda Zhao, Xueqian Wang, Gang Li, and You He. Frequency-adaptive learning for sar ship detection in clutter scenes. *IEEE TGARS*, 2023. 3, 6, 7

[43] Qiulin Zhang, Zhuqing Jiang, Qishuo Lu, Jia'nan Han, Zhengxin Zeng, Shang-Hua Gao, and Aidong Men. Split to be slim: An overlooked redundancy in vanilla convolution. *arXiv preprint arXiv:2006.12085*, 2020. 3

[44] Tianwen Zhang, Xiaoling Zhang, Jun Shi, and Shunjun Wei. Depthwise separable convolution neural network for high-speed sar ship detection. *Remote Sensing*, 11(21):2483, 2019. 1, 2

[45] Wenhua Zhang, Licheng Jiao, Fang Liu, Shuyuan Yang, Wei Song, and Jia Liu. Sparse feature clustering network for unsupervised sar image change detection. *IEEE TGARS*, 60: 1–13, 2022. 1

[46] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 3

[47] Yan Zhang, Ben-Yuan Gu, Bi-Zhen Dong, Guo-Zhen Yang, Hongwu Ren, Xueru Zhang, and Shutian Liu. Fractional gabor transform. *Optics letters*, 22(21):1583–1585, 1997. 4

[48] Yue Zhou, Xue Jiang, Guozheng Xu, Xue Yang, Xingzhao Liu, and Zhou Li. Pvt-sar: An arbitrarily oriented sar ship detector with pyramid vision transformer. *IEEE JSTARS*, 16: 291–305, 2022. 6