WAFormer: Ship Detection in SAR Images Based on Window-Aware Swin-Transformer

Zhicheng Wang[1,3], Lingfeng Wang[2](✉), Wuqi Wang[3], Shanshan Tian[3], and Zhiwei Zhang[3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

[2] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China
lfwang@mail.buct.edu.cn

[3] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract. The research work of synthetic aperture radar (SAR) image target detection based on deep learning has made great progress. However, most of them apply the methods applicable to optical images directly to SAR images, ignoring the characteristics of targets in SAR images. For instance, the size of target in SAR images is usually small and volatile. Meanwhile, the target distribution is relatively sparse and the detection is affected by the complex background noise. In this paper, we propose an improved backbone network, called WAFormer, for ship targets detection in SAR images, based on the latest Swin-Transformer. WAFormer improves the local window attention mechanism of Swin-Transformer by introducing the new window settings. Our model can be more suitable to match the shape of the target, so that it obtains more accurate detection in SAR images. Experimental results show that the WAFormer achieves 74.4% mAP on the Official-SSDD SAR dataset, surpassing Swin-Transformer by +1.0, especially for large targets.

Keywords: Synthetic-Aperture Radar (SAR) · Ship detection · Transformer · Window attention

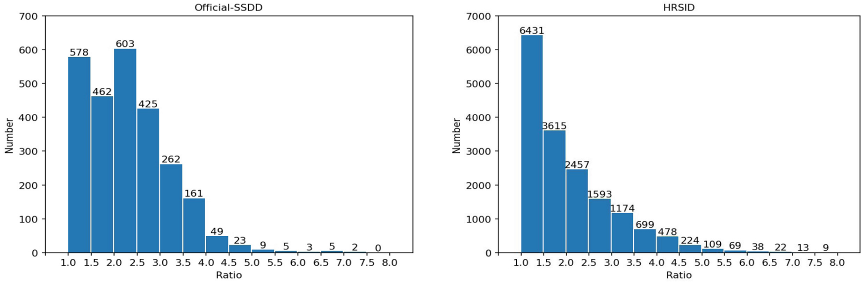
1 Introduction

As an active microwave remote sensing device, synthetic aperture radar (SAR) is capable generate all-day, all-weather and high-resolution earth observations. SAR images are of great importance in reconnaissance and surveillance missions in the military and civilian domains. SAR images target detection can be applied in many tasks, such as environmental monitoring, battlefield reconnaissance, geographic survey and ocean monitoring.

Deep learning technology has achieved excellent results in solving optical images detection and recognition tasks, and has attracted more and more scholars to use deep learning technology in SAR images interpretation tasks [1–4].

S. Yu et al. (Eds.): PRCV 2022, LNCS 13536, pp. 524–536, 2022.
https://doi.org/10.1007/978-3-031-18913-5_41

But the complex imaging mechanism of SAR images is different from optical image, leads to the fact that algorithms perform well on optical images may not be perfectly adapted to SAR images. In general, the challenges of applying deep learning to study the tasks of SAR images target detection are mainly as follows: (1) As shown in Table 1 and Fig. 1 statistics from Official-SSDD [5,6] and HRSID [7], two mainstream SAR image dataset, the size of sparse targets is generally small and the scale varies greatly, it undoubtedly increases the difficulty of SAR images target detection. (2) SAR images are often accompanied by cluttered noise and complex backgrounds such as docks, islands and reefs, resulting in lots of false detection or missed detection. (3) The difference between different datasets is large, lead to the generalization of the model trained on a single dataset is weak.



(a) Official-SSDD (b) HRSID

Fig. 1. Distribution of the ratio of the long side to the short side of the target bounding box.

As above, we aim to extract precise target features from complex SAR images to solve the problem of small target detection and multi-scale target detection. We propose an improved Transformer backbone based on Swin-Transformer [8] which called WAFormer. The backbone redesign the window attention module considering the size and shape of the SAR images targets. The improved window can better capture targets of various sizes and directions and distinguish them from the background. WAFormer achieves higher box AP than Swin-Transformer and other classic convolutional neural network (CNN) method with lower FLOPS than Swin-Transformer. Meanwhile we prove the Transformer method is suitable for SAR images target detection.

The main contributions of this paper are as follows:

(1) We redesign the Transformer window attention module with the size variable window. The resizable window make feature extraction more suitable for SAR images targets of various postures.
(2) To enhance connections between non-overlapping windows in abovementioned window attention module, we improve the original shift window mechanism in Swin-Transformer to make it more reasonable.

(3) In order to alleviate the computational redundancy problem caused by the new window attention, we introduce a channel splitting mechanism to calculate the window attention of different direction at the same time.

Table 1. Statistical results of multi-scale ships in Official-SSDD and HRSID.

Dataset	Size of ships (number)			Special size (pixels)	
	Small	Medium	Large	Smallest	Largest
Official-SSDD	1624	895	68	4 * 4	384 * 308
HRSID	9242	14776	321	3 * 1	800 * 653

2 Related Works

2.1 SAR Target Detection Based on Deep Learning

The analysis of SAR images data has become a research hot spot because of its significance in the field of military and civil detection. In recent years, many SAR images target detection methods based on deep learning are gradually developed. Cui et al. [9] utilized a dense attention pyramid network (DAPN) to improve the accuracy of multi-scale ship detection. Zhao et al. [10] proposed an attention receptive pyramid network (ARPN) with receptive fields block (RFB) and convolutional block attention module (CBAM) to improve the performance of detecting multi-scale ships. Cui et al. [11] proposed an anchor-free method which introduces spatial shuffle-group enhance (SSE) attention module to CenterNet to achieve better performance than some classic CNN methods. Fu et al. [12] are also based on anchor-free strategy, proposed a feature balancing and refinement network (FBR-Net) to achieve the state-of-the-art performance among the general anchor-free methods. Guo et al. [13] presented CenterNet++ consists of feature refinement module, feature pyramids fusion module, and head enhancement module to improve the effectiveness and robustness. Tang et al. [14] proposed a scale-aware feature pyramid network comprises a scale-adaptive feature extraction module and a learnable anchor assignment strategy to address the problem of feature misalignment and targets' appearance variation. Xu et al. [15] improved YOLOv5 to present Lite-YOLOv5, a lightweight onboard SAR ship detector with decreasing FLOPS and without sacrificing accuracy. Xia et al. [16] proposed a visual transformer framework based on contextual joint-representation learning by combining the global information of Transformer and the local feature representation of CNN.

2.2 Vision Transformer

Transformer [17] is the framework of encoder-decoder with attention mechanism for natural language processing (NLP). With Transformer's impressive performance in NLP, a growing number of computer vision research work based on Transformer has emerged. ViT [18] presented a pure Transformer architecture for vision by inputting the patches sequences splitted from an image to Transformer. But when the training data is not sufficient ViT will not generalize well. Also based on convolution-free Transformers, DeiT [19] introduced distillation strategy into Transformer to achieve competitive performance. DEtection TRansformer (DETR) [20] realized an end to end detector including a transformer encoder-decoder architecture and a global loss calculated in the parallel decoder. PVT [21] introduced pyramid structure to Transformer to generate an excellent vision Transformer backbone with lower computation than ViT. But these methods based on global attention have high computational complexity. Swin-Transformer [8] presented a general vision Transformer backbone which innovatively designed the shifted windows based on hierarchical architecture. The non-overlapping local windows attention mechanism and cross-window connection not only reduces the computational complexity, but also realizes the state-of-the-art of multiple visual tasks. CSwin [22] proposed a cross-shaped window consists of horizontal and vertical stripes split from feature in a parallel manner, meanwhile introduced Locally-enhanced Positional Encoding (LePE) to achieve better position encoding ability. However, local window attention is not friendly to big target detection. Our method optimizes this disadvantage inspired by Swin-Transformer and CSwin to optimize this disadvantage.

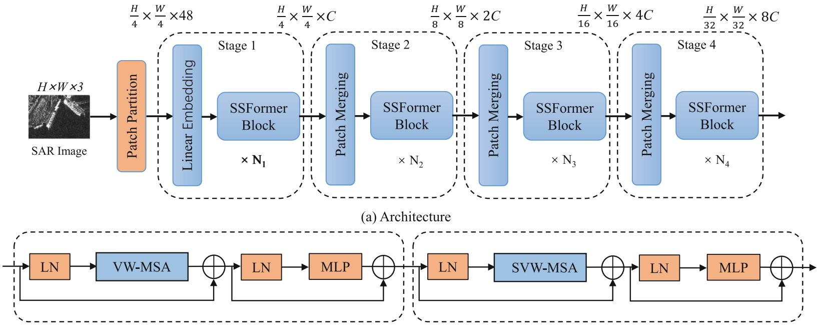
3 Method

3.1 Motivation

Swin-Transformer [8] is currently state-of-the-art vision Transformer backbone with higher accuracy and lower cost than others. The excellent feature extraction capability and advantages for small target detection of the window attention mechanism inspired us to apply it to SAR images target detection. Nevertheless, due to characteristics of small and diverse target size, sparse distribution and different postures, Swin-Transformer can not be directly applied to SAR images. Thus we redesign the window with variable size and apply it to the original Swin structure, formed the improved backbone for ship target detection in SAR images, called WAFormer.

3.2 Overview

The overall architecture of WAFormer is shown in Fig. 2. Because the proposed method is based on Swin-Transformer, so that the overall structure of the network tends to be similar. Taking an image as input, same to Swin-Transformer, followed with the patch partition module to split the image into evenly divided patches. Then applying a linear embedding layer project the patch tokens to C dimension. The setting of patch size and the number of tokens, and the design of the hierarchical representation are both same to Swin-Transformer, so that we also have $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$ tokens in the i^{th} stage with decreased resolution and increased channels. The difference is that we replace the original Swin-Transformer block with our WAFormer block. The WAFormer block will be described in detail as follows.



(a) Architecture

(b) WAFormer Block

Fig. 2. (a) The overall architecture of our proposed WAFormer; (b) an effective Transformer block for ship detection in SAR images described in Sect. 3.4). VW-MSA and SVW-MSA are multi-head attention modules with vertical/horizontal and shifted windowing configurations, respectively.

3.3 Variable Size Window Self-attention

Variable Size Window. Based on the local window attention mechanism, we propose a variable size window more suitable for ship target in SAR images. Firstly, in order to allow multi-scale input, the image is padded. Then the padded feature is partitioned into non-overlapping windows. The window size is set as $M \times N$ mean that each window contains $M \times N$ patches. As shown in Fig. 1, statistics indicate that the ratio of long and short sides of the bounding box of SAR images is mostly in the range of 4:1. While the aspect ratio of the window of Swin-Transformer is 1:1 which can not cover all targets and will truncate some targets. Thus we set the window size according to this ratio range as shown in Fig. 3. Specifically, from "Stage 1" to "Stage 4", we empirically set the long and short sides of the window to $\frac{224}{7*2^{i-1}}$(i = 1, 2, 3, 4) and [7, 4, 2, 1]. Meanwhile, we set horizontal and vertical windows to capture the targets of different postures. Inspired by CSWin [22], we introduce the channel split method to calculate horizontal and vertical window attention at the same time to reduce costs.

Shifted Window. Since our window is no longer a fixed size, the original shifted window is not applicable. To increase the connection between non-overlapping windows, we replace the original shift step with $\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$ to displace the regularly partitioned windows. In other words, the shift size becomes half of the short side of the window, which is proved to be effective by experiments.

Convolution Position Encoding. It is well known that position encoding is of great significance to the Transformer model [17,26,27]. However, we abandoned absolute position encoding and chose to utilize relative position encoding. Because we notice that the absolute position encoding does not lead to performance improvement. Inspired by LePE of CSWin, we also utilize a learnable additive positional encoding by performing convolution operation on *value* V of the window. We calculate the attention for a window according to the following formula:

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d})V + Conv(V) \tag{1}$$

Experiments show that this position encoding can effectively improve the accuracy.

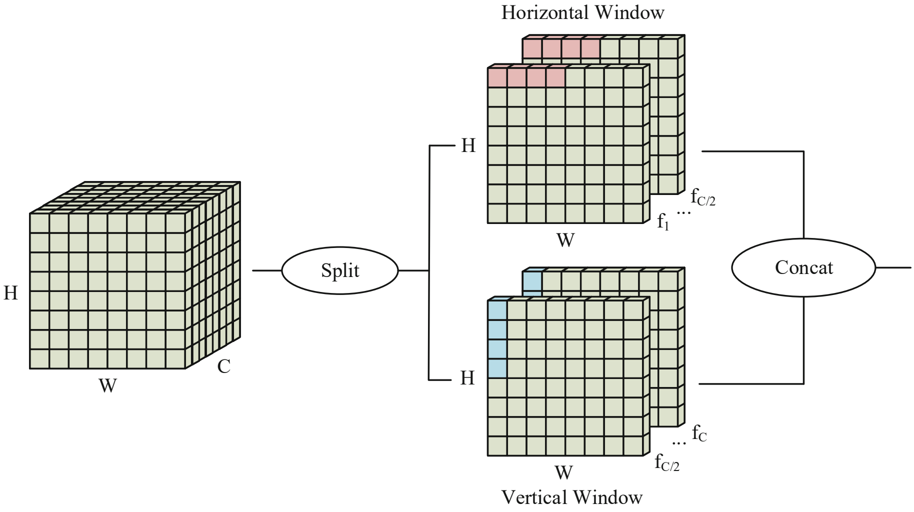


Fig. 3. The illustration of the variable size window with channel splitting manner

Computation Complexity Analysis. Omitting SoftMax, the computation complexity of a variable size window attention module based on an SAR image of $h \times w$ patches is:

$$\Omega(VW\text{-MSA}) = 4hwC^2 + 2MNhwC \tag{2}$$

where *hw* denote the patch num, it can be seen that our computational complexity is also linear with hw when *MN* set as we design.

3.4 WAFormer Block

Our network is built on WAFormer block, with other layers kept same with Swin-Transformer. A WAFormer block contains a pair of regular and shifted variable size window attention modules. This block is defined as:

$$\begin{aligned}
\hat{X}^l &= VW\text{-}MSA(LN(X^{l-1})) + X^{l-1}, \\
X^l &= MLP(LN(\hat{X}^l) + \hat{X}^l), \\
\hat{X}^{l+1} &= SVW\text{-}MSA(LN(X^l)) + X^l, \\
X^{l+1} &= MLP(LN(\hat{X}^{l+1}) + \hat{X}^{l+1}),
\end{aligned} \tag{3}$$

where VW - MSA and SVW - MSA respectively denote the regular and shifted variable size window attention modules; $\hat{X}^l$ and X^{l+1} denote the output feature of the (S)VW - MSA module and the MLP module for block *l*.

4 Experiments

4.1 Dataset and Evaluation Metrics

SSDD [6] is the first open dataset which is widely used in the SAR remote sensing community. It includes 1160 SAR images with about 500 × 500 pixels and under 1–15 m resolutions. The dataset contains 2456 ship targets of different sizes and materials, good and bad sea condition, offshore and inshore scenes. Official-SSDD [5] is an optimized version based on the initial SSDD. Compared to SSDD, Official-SSDD revises labels, formulates stricter using standards and provides a comprehensive data analysis. HRSID [7] includes 5604 SAR images with 800×800 pixels and three resolutions(0.5 m, 1 m, 3 m). It contains 16951 ship targets covering different resolutions, polarization, sea condition, sea area, coastal port. We choose Official-SSDD as the main training and testing dataset, and HRSID as the validation dataset for comparison with Swin-Transformer. For detection evaluation metrics, we apply the mean Average Precision (mAP), detection rate at IOU = 0.5 (AP_{50}) and IOU = 0.75 (AP_{75}), and detection performance of target detection on small, medium, large targets (AP_S, AP_M, AP_L). The FLOPS and parameters of model used are also calculated and compared.

4.2 Implementation Details

We implement our proposed network on the PyTorch framework and MMDetection [23] toolbox. Multi-scale training [20,24] and data augmentation techniques [19] are adopted while the largest size is set as 1333 × 800 refer to Swin-Transformer. The experiments run at a NVIDIA GeForce RTX 3090 GPU and the batch size is set as 4 limited by the compute capability. The initial learning rate and training epoch are set as 0.0001 and 300. We use AdamW [25] optimizer and cosine decay learning rate scheduler with 5 epochs of linear warm-up. The weight decay is set as 0.05.

4.3 Comparison Results

We compare our proposed WAFormer backbone with Swin-Transformer using Mask R-CNN [28] object detection framework. Meanwhile, we also choose 5 classic object detection methods including YOLOv3 [29], SSD-512 [30], RetinaNet [31], Faster R-CNN [32], Mask R-CNN using ResNet-50 [33] as backbone. Figure 4 shows the visual results on Official-SSDD of WAFormer and Swin-Transformer with Mask R-CNN framework compared with other classic methods. It can be seen that the detection performance of our method is better than Swin-Transformer, and the confidence of the detection box is higher than that of other methods.

Table 2. Detection results on Official-SSDD test set.

Method	Image size	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv3	1024^2	65.7	96.1	78.1	66.3	65.3	67.7
SSD-512	512^2	70.1	96.3	84.4	70.1	71.1	74.4
RetinaNet	1024^2	73.8	98.3	88.8	73.2	76.4	**80.3**
Faster R-CNN	1024^2	73.1	96.7	88.1	71.6	**78.2**	76.9
Mask R-CNN R-50	1024^2	73.5	96.8	87.8	72.0	**78.2**	75.0
Mask R-CNN Swin	1024^2	73.4	97.7	89.6	73.3	74.7	62.9
Mask R-CNN WAFormer	1024^2	**74.4**	**98.6**	**90.4**	**73.7**	77.9	71.8

Table 3. Parameter size and FLOPs of methods in experiment.

Method	Image size	#Params	FLOPs
YOLOv3	1024^2	61.52M	198.5G
SSD-512	512^2	24.39M	87.12G
RetinaNet	1024^2	36.1M	209.13G
Faster R-CNN	1024^2	41.12M	211.28G
Mask R-CNN R-50	1024^2	43.75M	262.76G
Mask R-CNN Swin	1024^2	47.37M	267.01G
Mask R-CNN WAFormer	1024^2	41.31M	250.55G

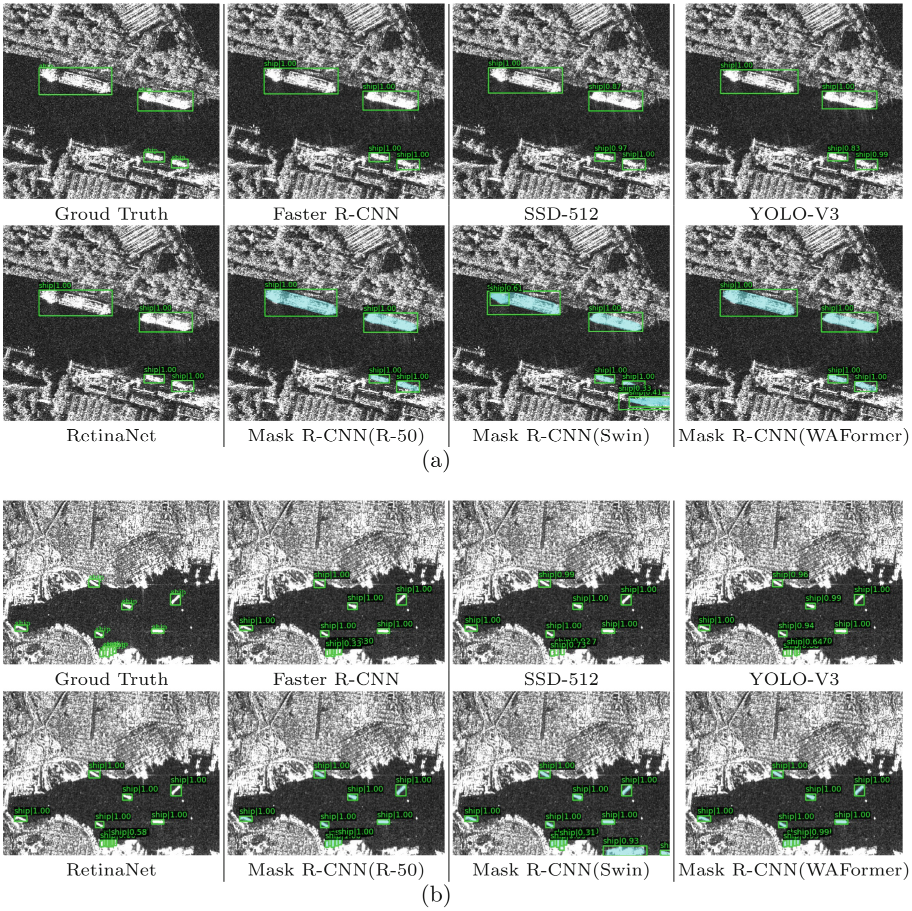


Fig. 4. Visual results of methods involved on Official-SSDD. R-50 namely ResNet-50 and Swin namely Swin-Tranformer.

Table 2 shows the performance comparisons of WAFormer with Swin-Transformer and other methods. Our WAFormer architecture achieves the highest detection accuracy among all the methods involved in the comparison. Specifically, our method achieves 74.4% mAP surpassing Swin-Transformer by +1.0, while the AP_{50} and AP_{75} are also bring advantages of +0.9 and +0.8 respectively. Meanwhile, we achieve the best result at AP_S and competitive result at AP_M with 73.7% and 77.9% respectively. The results demonstrate that our method brings improvements for solving small and multi-scale targets detection of SAR images. Table 3 shows the parameters and FLOPs of these methods. When using Mask R-CNN detection framework, our WAFormer realize less parameters and FLOPs than Swin-Transformer. Our method achieves the best results with a lighter architecture. This further shows the effectiveness and superiority of WAFormer for target detection in SAR images.

To validate the universality of our method over Swin-Transformer in SAR images target detection, we retrain and test WAFormer and Swin-Transformer with Mask R-CNN framework on HRSID. Table 4 shows that we still have advantage compared with Swin-Transformer.

Table 4. Detection results on HRSID test set.

Method	Image size	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN Swin	1024^2	64.3	87.0	75.3	65.3	67.4	38.5
Mask R-CNN WAFormer	1024^2	**65.1**	**87.2**	**75.5**	**65.8**	**68.2**	**44.2**

4.4 Related Configuration Adjustment

Window Size and Shift Size. To achieve the optimal performance, we conducted different configuration experiments on the size and the shift size of the window. Table 5 shows the results of different configuration. The results show that the highest accuracy is achieved when the long side and short side are of the window set as [32, 16, 8, 4] and [7, 4, 2, 1]. And when the shift size is set as $\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$, the shifted window can bring optimal performance.

Table 5. The performance of different configuration on size of the window and step size of the shifted window. The long side and short side denote the size of the window.

Long side	Short side	Shift size	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
[56, 28, 14, 7]	[8, 4, 2, 1]	$\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$	73.4	97.7	90.5	73.3	74.6	71.0
[56, 28, 14, 7]	[7, 4, 2, 1]	$\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$	73.5	97.7	89.3	73.0	76.5	68.8
[32, 16, 8, 4]	[8, 4, 2, 1]	$\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$	73.9	98.6	89.1	73.4	76.6	**77.7**
[32, 16, 8, 4]	[7, 4, 2, 1]	$\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$	**74.4**	**98.7**	90.4	**73.8**	77.4	71.0
[16, 8, 4, 2]	[8, 4, 2, 1]	$\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$	73.7	98.5	**90.7**	73.2	75.8	70.0
[16, 8, 4, 2]	[7, 4, 2, 1]	$\left(\left\lfloor \frac{short\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$	73.3	97.7	89.8	72.6	76.1	65.9
[32, 16, 8, 4]	[7, 4, 2, 1]	$\left(\left\lfloor \frac{long\text{-}side}{2} \right\rfloor, \left\lfloor \frac{short\text{-}side}{2} \right\rfloor\right)$	73.6	98.6	90.4	72.7	76.8	73.8
[32, 16, 8, 4]	[7, 4, 2, 1]	$\left(\left\lfloor \frac{long\text{-}side}{2} \right\rfloor, \left\lfloor \frac{long\text{-}side}{2} \right\rfloor\right)$	73.9	98.6	89.4	73.0	**77.5**	66.7

Convolution Position Encoding. To validate the effect of convolutional relative position encoding, we also conducted relevant experiments. We calculate the origin attention without the convolution position encoding, the attention with additive and multiplicative convolutional position encoding, respectively. The results show in Table 6, the results show that the additive convolutional position encoding is beneficial to improve the accuracy.

Table 6. The performance of different position encoding. mul conv rel pos.: multiplicative convolutional position encoding, add conv rel pos.: additive convolutional position encoding

Position encoding	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
No pos.	73.3	97.8	90.3	72.5	76.5	71.3
Mul conv rel pos.	73.6	**98.6**	89.4	72.5	**77.9**	**77.7**
Add conv rel pos.	**74.4**	**98.6**	**90.4**	**73.7**	**77.9**	71.8

5 Conclusion

In this paper, according to the characteristics of the SAR images, we propose a backbone focus on target size based on Swin-Transformer. Our method improves the target detection performance in SAR images while reducing the cost. Experiments show the targeted improvements have played an effective role in solving the problem of difficult detection of small and multi-scale targets in SAR images. At the same time, our size variable window is also applicable to other datasets, since it is designed according to the dataset. However, it can be found that our large target detection results are not excellent. We consider this may be a shortcoming of window attention mechanism. In future work, we plan to increase the number of large windows in the shallow layer, and introduce the channel attention mechanism to increase the information interaction between channels.

References

1. Jiao, J., et al.: A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection. IEEE Access **6**, 20881–20892 (2018)
2. Chang, Y.-L., Anagaw, A., Chang, L., Wang, Y.C., Hsiao, C.-Y., Lee, W.-H.: Ship detection based on YOLOv2 for SAR imagery. Remote Sens. **11**(7), 786 (2019)
3. Zhang, T., Zhang, X.: High-speed ship detection in SAR images based on a grid convolutional neural network. Remote Sens. **11**(10), 1206 (2019)
4. An, Q., Pan, Z., Liu, L., You, H.: DRBox-v2: an improved detector with rotatable boxes for target detection in SAR images. IEEE Trans. Geosci. Remote Sens. **57**(11), 8333–8349 (2019)
5. Zhang, T., et al.: SAR ship detection dataset (SSDD): official release and comprehensive data analysis. Remote Sens. **13**(18), 3690 (2021)
6. Li, J., Qu, C., Shao, J.: Ship detection in SAR images based on an improved faster R-CNN. In: 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), pp. 1–6. IEEE (2017)
7. Wei, S., Zeng, X., Qu, Q., Wang, M., Su, H., Shi, J.: HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation. IEEE Access **8**, 120234–120254 (2020)
8. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

9. 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. **57**(11), 8983–8997 (2019)
10. Zhao, Y., Zhao, L., Xiong, B., Kuang, G.: Attention receptive pyramid network for ship detection in SAR images. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. **13**, 2738–2756 (2020)
11. 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. **59**(1), 379–391 (2020)
12. 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. **59**(2), 1331–1344 (2020)
13. Guo, H., Yang, X., Wang, N., Gao, X.: A CenterNet++ model for ship detection in SAR images. Pattern Recogn. **112**, 107787 (2021)
14. Tang, L., Tang, W., Qu, X., Han, Y., Wang, W., Zhao, B.: A scale-aware pyramid network for multi-scale object detection in SAR images. Remote Sens. **14**(4), 973 (2022)
15. Xu, X., Zhang, X., Zhang, T.: Lite-YOLOv5: a lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 SAR images. Remote Sens. **14**(4), 1018 (2022)
16. Xia, R., et al.: CRTransSar: a visual transformer based on contextual joint representation learning for SAR ship detection. Remote Sens. **14**(6), 1488 (2022)
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
19. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
20. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
21. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
22. Dong, X., et al.: CSWin transformer: A general vision transformer backbone with cross-shaped windows. arXiv preprint arXiv:2107.00652 (2021)
23. Chen, K., et al.: MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
24. Sun, P., et al.: Sparse R-CNN: end-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14454–14463 (2021)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning, pp. 1243–1252. PMLR (2017)
27. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)

28. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
29. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
30. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
31. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
32. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
33. 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, pp. 770–778 (2016)