remote sensing

MDPI

Communication

SAR Target Detection Based on Improved SSD with Saliency Map and Residual Network

Fang Zhou [1], Fengjie He [1], Changchun Gui [1], Zhangyu Dong [1] and Mengdao Xing [1,2,*]

[1] School of Computer and Information, Hefei University of Technology, Hefei 230009, China; zhoufang@hfut.edu.cn (F.Z.); 2019111044@mail.hfut.edu.cn (F.H.); 2020111048@mail.hfut.edu.cn (C.G.); dzyhfut@hfut.edu.cn (Z.D.)

[2] Institute of Electronic Engineering, Xidian University, Xi'an 710071, China

* Correspondence: xmd@xidian.edu.cn

Abstract: A target detection method based on an improved single shot multibox detector (SSD) is proposed to solve insufficient training samples for synthetic aperture radar (SAR) target detection. We propose two strategies to improve the SSD: model structure optimization and small sample augmentation. For model structure optimization, the first approach is to extract deep features of the target with residual networks instead of with VGGNet. Then, the aspect ratios of the default boxes are redesigned to match the different targets' sizes. For small sample augmentation, besides the routine image processing methods, such as rotating, translating, and mirroring, enough training samples are obtained based on the saliency map theory in machine vision. Lastly, a simulated SAR image dataset called Geometric Objects (GO) is constructed, which contains dihedral angles, surface plates and cylinders. The experimental results on the GO-simulated image dataset and the MSTAR real image dataset demonstrate that the proposed method has better performance in SAR target detection than other detection methods.

Keywords: synthetic aperture radar (SAR); target detection; deep learning; saliency map

Citation: Zhou, F.; He, F.; Gui, C.; Dong, Z.; Xing, M. SAR Target Detection Based on Improved SSD with Saliency Map and Residual Network. *Remote Sens.* **2022**, *14*, 180. https://doi.org/10.3390/rs14010180

Academic Editors: Mi Wang, Hanwen Yu, Jianlai Chen and Ying Zhu

Received: 29 October 2021
Accepted: 28 December 2021
Published: 1 January 2022

1. Introduction

Synthetic aperture radar (SAR) is an active earth observation system with high resolution. The scene imaging quality is considerable with optical images; these images are increasingly important for scientific applications [1–4]. With the increasing improvement of SAR data collection capability and imaging algorithms, the research on interpreting high-resolution SAR images has received extensive attention, such as target detection and change detection [5,6]. Traditional SAR target detection algorithms include the constant false alarm rate (CFAR) method [7,8], template matching method [9,10], etc. These methods primarily embark on feature extraction and classifier design, which require highly manual involvement, complex design process, and have poor detection performance in complex scenes.

A convolutional neural network (CNN) has advantages of high recognition accuracy and generalization ability. It can also actively extract features without manual design. In addition, CNN performs fully supervised learning based on labeled information. Since the R-CNN [11] model first introduced the convolutional neural network into the field of target detection, many target detection algorithms with excellent performance have been proposed successively. These algorithms are usually divided into two categories: two-stage detection algorithms and single-stage detection algorithms. The two-stage detection algorithm has higher detection accuracy but slower detection speed. Representative algorithms include R-CNN, Fast R-CNN [12], Faster R-CNN [13], and R-FCN [14]. The single-stage detection algorithm gets the category information and position directly by regression, taking into account the accuracy and velocity. Representative algorithms are SSD [15], YOLO [16–19], Retina-net [20], etc.

CNN has been widely studied as typical deep learning models in SAR target detection. Ding et al. [21] proposed three data enhancement methods for training samples to recognize SAR targets using CNN, and they also validated the effectiveness of data enhancement and good detection capability for SAR images with random speckle noise through experiments. Fei et al. [22] developed three improvement techniques to enhance the feature extraction ability of the ship detection network, which contains multi-level sparse optimization of SAR image, a novel split convolution block (SCB) and a spatial attention block (SAB). Lin et al. [23] proposed a highway neural network for limited labeled SAR training data, consisting of a modified convolutional highway layer, a maximum pooling layer, and a dropout layer. Deeper feature information can be extracted from limited data. For MSTAR target classification, the model can achieve 94.97% recognition accuracy, with only 30% of the original training samples, and 99% recognition accuracy when a complete training set is used. He et al. [24] proposed a SAR target recognition model with multi-angle tensor sparse representation by combining the local structural features of the target and the correlation between multiple SAR images of the same target. Jun et al. [25] proposed a hierarchical convolutional neural network (H-CNN), which is a two-stage CNN model. They extracted regions of interest (ROIs) in the coarse-detection stage and used CNN in the fine-detection stage. The model is better than the conventional constant false-alarm rate and CNN-based models. Du et al. [26] proposed a three-channel sub-aperture synthesis algorithm to transfer the pre-trained network weights on optical images to SAR images. In comparison with two-parameter CFAR and Faster R-CNN, the model has better detection performance. Zhang et al. [27] proposed a novel two-phase object-based deep learning approach for SAR image change detection, which includes object-based approach, superpixel objects and two-phase deep learning framework, significantly reducing false alarm rates, leading to 99.71% change detection accuracy. Hao Chen et al. [28] proposed a spatial–temporal attention neural network (BAM and PAM) for remote sensing image binary CD, which mitigates misdetection caused by misregistration in bitemporal images, showing better robustness in color and scale variations.

The limitation of training samples for SAR target detection leads to some difficulties, such as overfitting, gradient dispersion/explosion, and network degradation. Convolutional neural networks are generally optimized by a gradient-based backpropagation (BP) algorithm. For feedforward networks, it is necessary to propagate the input forward, propagate the error backward, and use the gradient method to update the parameters. The parameter update of the k-th layer needs to calculate the gradient of the loss, which depends on the error term of the layer. According to the chain rule, the error term of the k-th layer depends on the error term of the $k+1$-th layer. In deep networks, since the size of k-th layer error term cannot be guaranteed, gradient dispersion/explosion easily happens. On the premise that the convolutional neural network can converge, as the depth of the network increases, the performance of the network first gradually increases to saturation, and then rapidly decreases [29].

We set out to solve these problems from two aspects. One is optimizing the model structure: we improve the backbone network by deeper residual network and redesign ratios of the default boxes to match the different targets' sizes. Compared to SSD original feature extraction, the residual network can be implemented in the form of skip connection, i.e., the input of the unit is directly added to the output of the unit and finally activated. Hence, the residual network can directly use the BP algorithm to update the parameters. The advantages of the SSD are obvious: the running speed is comparable to that of YOLO, and the detection accuracy is comparable to that of Faster R-CNN. Although the idea of pyramdial feature hierarchy is adopted, it is not good for small target detection. This is because the feature map extracted by the SSD in the shallow layer is not strong enough. As the depth deepens, the information of the small target in the high-level feature map is easy to lose. We use the residual network to improve the backbone and increase the size of the input picture from 300 × 300 × 3 to 600 × 600 × 3. The second method is saliency map augmentation: we apply a bilinear interpolation saliency mapping method to build an

image pyramid model and use the model to process all of the training samples. Combined with routine image processing methods, we obtain sufficient training samples to improve the recognition accuracy. Our proposed method shows good detection performance based on both the real SAR data and the simulated SAR data.

2. Methods

2.1. Feature Extraction

The structure diagram of the improved SSD detection model is given in Figure 1. The original SSD model uses a pre-trained VGG16 as the backbone, which has limitations for image feature extraction due to the shallow layers. The performance of a deep learning model can be improved by deepening the network, but its accuracy is not linearly related to the depth of the network. As the neural network is going deeper, the model classification accuracy gradually rises to saturation, then decreases. At the same time, deeper networks imply more weight parameters, which tend to cause overfitting when the training samples are small. In this paper, we improve the backbone by residual network, and the specific network structure is shown in Figure 2, with a 7×7 convolutional layer, a 3×3 pooling layer, and three densely connected residual blocks from the left to right. Each residual block contains varying amounts of basic residual network units, and the numbers of network units are 3, 4, and 6, respectively. The improved backbone network structure parameters are given in Table 1.

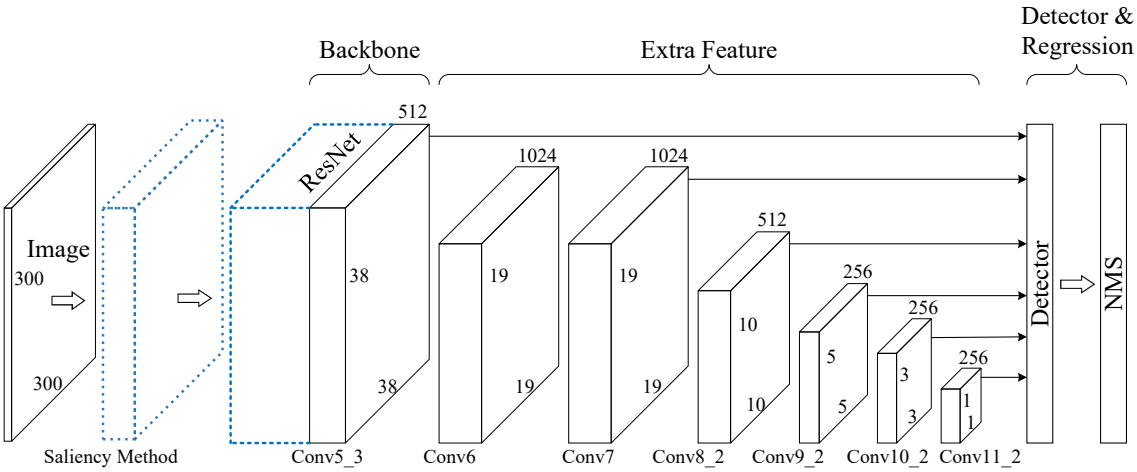


Figure 1. Flowchart of the proposed SAR target detection method.

Table 1. ResNet feature extraction structure parameters.

Layer Name	Parameters	Output Feature Size
Convolutional layer	7×7 Conv, stride 2	$300 \times 300 \times 64$
Pooling layer	3×3 Max pool, stride 2	$150 \times 150 \times 64$
Residual block 1	$\begin{bmatrix} 1 \times 1 \text{ Conv} \\ 3 \times 3 \text{ Conv} \\ 1 \times 1 \text{ Conv} \end{bmatrix} \times 3$	$150 \times 150 \times 256$
Residual block 2	$\begin{bmatrix} 1 \times 1 \text{ Conv} \\ 3 \times 3 \text{ Conv} \\ 1 \times 1 \text{ Conv} \end{bmatrix} \times 4$	$75 \times 75 \times 512$
Residual block 3	$\begin{bmatrix} 1 \times 1 \text{ Conv} \\ 3 \times 3 \text{ Conv} \\ 1 \times 1 \text{ Conv} \end{bmatrix} \times 6$	$38 \times 38 \times 512$

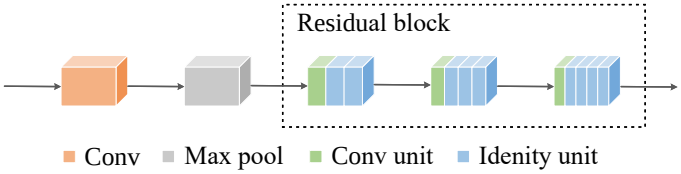


Figure 2. ResNet feature extraction structure.

As shown in Figure 3, there are two types of basic residual network units: Conv Unit and Identity Unit. As can be seen from Table 1, there is a 1×1 convolutional layer before and after the 3×3 convolutional layer in basic residual network units. The 1×1 convolutional layers are responsible for reducing and increasing the dimensions so that the 3×3 convolutional layer has smaller input/output dimensions. By using the bottleneck architectures, the input and output feature channels can be kept consistent, and the computational work can be significantly reduced. At the same time, the skip connections make the input of the units directly affect the output, so it can alleviate the gradient vanishing. As shown in Figure 3, Batch Normalization (BN) processing is added to both basic units. To optimize the convolutional neural network, BN adjusts the data distribution of the output of the previous convolutional layer in the form of a zero-mean and one-variance distribution. This ensures the validity of the gradient and accelerates the convergence of the model.

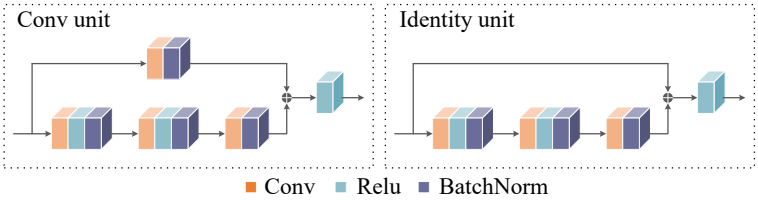


Figure 3. The basic residual network units.

2.2. Small Sample Augmentation

To solve the problem of insufficient real SAR data, traditional small sample augmentation includes translation, rotation, cropping, etc. In addition to these methods, we present the saliency map method to obtain sufficient training samples. A saliency map is an image that shows the characteristic of each pixel. Pixels with higher gray levels in RGB images are displayed in a more distinctive way in the saliency map. Its purpose is to transform an image into a more analyzable form.

As shown in Figure 4, firstly, we construct a Gaussian pyramid model with image intensity components. Then, the model is upsampled to the same size as the original image to obtain images of different resolutions. Finally, the saliency map is made by the sum of the difference between the different levels of the model. The image M is downsampled $j (j = 0, \ldots, M)$ times to get an image M_j, which is $1/2^j$ of the original size, and M_0 represents the original image. The coordinates of the four pixel points Q_{11}, Q_{12}, Q_{21} and Q_{22} in the image M_j are known as (x_1, y_1), (x_1, y_2), (x_2, y_1) and (x_2, y_2), respectively, and the coordinate of the point P to be interpolated is (x, y). The interpolation is performed first along the x-axis to obtain the intensity values of the temporary pixel points $A(x, y_1)$ and $B(x, y_2)$:

$$\begin{aligned} f(A) &\approx \frac{x - x_1}{x_2 - x_1} f(Q_{21}) + \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) \\ f(B) &\approx \frac{x - x_1}{x_2 - x_1} f(Q_{22}) + \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) \end{aligned} \tag{1}$$

where $f(\cdot)$ represents the pixel intensity value. Then, along the y-axis to obtain the ultimate pixel point intensity value:

$$f(P) \approx \frac{y_2 - y}{y_2 - y_1} f(A) + \frac{y - y_1}{y_2 - y_1} f(B) \tag{2}$$

when moving from the bottom to the top of the Gaussian pyramid, the resolution and size of the image are reduced. The bottom of the image pyramid is a high-resolution representation of the image M, which retains more detailed information, while the top is a low-resolution representation of the image M, which retains more background information. Thus, we upsample the smaller size image to the same size as the original image to obtain $M_j^{'}(j = 0, \ldots, 7)$ by bilinear interpolation. The difference results of the target and the surrounding background at different levels are obtained by making the pairwise difference operation. The difference operation we used here is $\{M_0^{'} - M_1^{'}, M_0^{'} - M_3^{'}, M_0^{'} - M_5^{'}, M_4^{'} - M_1^{'}, M_2^{'} - M_3^{'}, M_3^{'} - M_5^{'}\}$. Then, we add the absolute values of the difference results together and regularize the summation to the range of (0, 255) to obtain the saliency map of the original image. Figure 4 illustrates the saliency map generation process, and Figure 5 shows an example of the saliency map results.

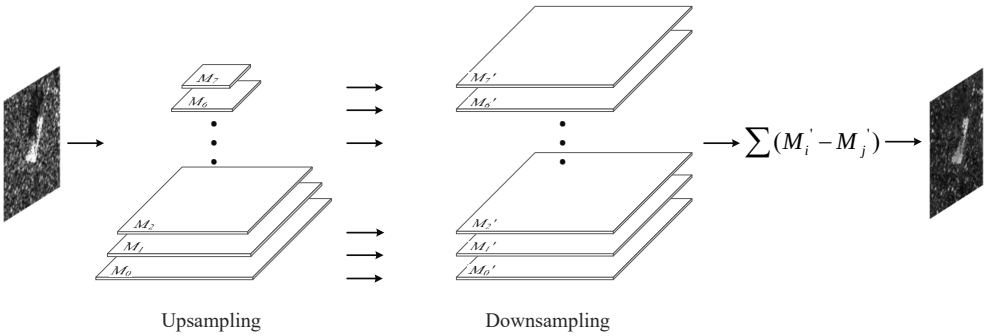
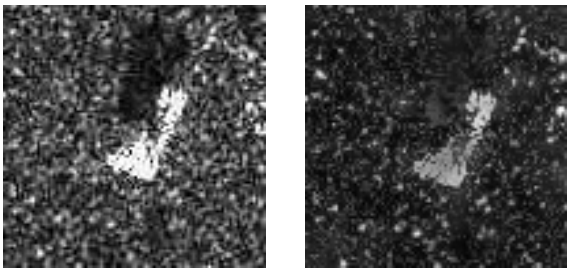


Figure 4. The saliency map generation process of an image.



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

Figure 5. Saliency map result. (**a**) 2S1, (**b**) saliency map of 2S1 after transforming.

2.3. Aspect Ratios of Default Boxes

In this paper, we use different scale feature maps to match varying sizes of objects in an image. For each feature map, default boxes are generated at different scales and ratios (e.g., 3×3 and 5×5 in Figure 6a), and the predicted bounding boxes are based on these default boxes. The image can be divided into more grids by the 5×5 feature map, but the default boxes of these grids are smaller than those in the 3×3 feature map, as shown in Figure 6b. So the 5×5 feature map can be used to detect small target and the 3×3 feature map can be used to detect the larger one relatively, as shown in Figure 6c.

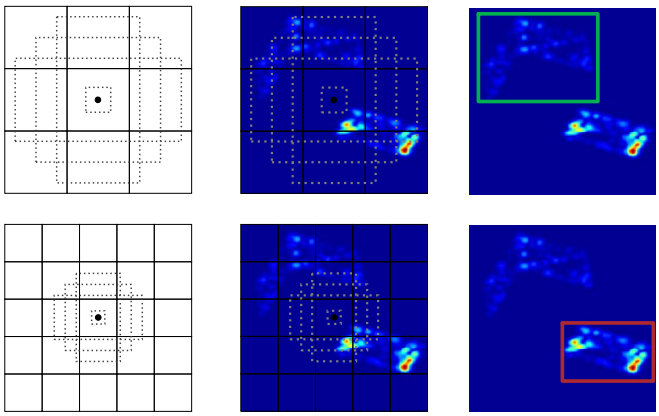
The center coordinates of the default box are the center of each grid $(\frac{a+0.5}{|f_k|}, \frac{b+0.5}{|f_k|})$, where $|f_k|$ is the k-th layer feature map size, a,$b \in \{0, 1, \ldots, |f_k|\}$. Each default box scale is computed as:

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{n - 1}(k - 1), k \in [1, n] \quad (3)$$

where n is 6, s_{max} is 0.9 and s_{min} is 0.2, meaning the highest layer feature map has a scale of 0.2 and the lowest layer feature map has a scale of 0.9. The width and height of the default boxes are computed based on the scale and ratio:

$$\begin{aligned} w_k^m &= s_k\sqrt{r_m}, \quad m \in \{1,2,3,4,5\} \\ h_k^m &= s_k/\sqrt{r_m}, \quad m \in \{1,2,3,4,5\} \end{aligned} \tag{4}$$

where w_k^m and h_k^m are the width and height of the m-th default box of the k-th layer feature map, and aspect ratios are $r_m \in \{1, 2, 3, 1/2, 1/3\}$. When $r_m = 1$, the scale of default box is $s_k^{'} = \sqrt{s_k s_{k+1}}$, $w_k^6 = h_k^6 = \sqrt{s_k s_{k+1}}$. In practice, we can adjust the aspect ratios to match targets in a specific dataset better. Our dataset includes dihedral angles, surface plates and cylinders. The cylinder's aspect ratio is much smaller than 1/3, so we set the aspect ratios as $r_m \in \{1, 2, 6, 1/2, 1/6\}$ in training.



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

Figure 6. Different scale detection. (**a**) Different scale feature maps: 3 × 3 and 5 × 5, (**b**) prediction procedure, (**c**) detection result.

3. Experimental Results and Discussion

3.1. Datasets

We carried out experiments on two different datasets: the Geometric Objects (GO) dataset and the MSTAR dataset. The GO dataset is an electromagnetic simulation SAR image dataset, including three geometries: dihedral angles, surface plates, and cylinders. For each geometry, we collect different azimuth and elevation simulation data, ranging from −6.3° to 96.4° for azimuth and 10° to 90° for elevation. The MSTAR dataset is a representative public dataset for SAR target recognition, which includes a total of 10 categories of military targets. The resolution for MSTAR images is 0.3 m × 0.3 m, with azimuth from 0° to 360°, and elevation of 15° and 17°. Here, we select three types of military targets: 2S1, D7 and T62, for training and testing. Training set and testing set are divided as shown in Table 2. We only carry out small sample augmentation on the training set.

Table 2. The quantity of training images and testing images.

Dataset	Training Set		Testing Set
	Initial Quantity	After Augmentation	
GO	131	3930	230
MSTAR	821	1642	897

3.2. Training Strategy

We verified the effectiveness of the improved method through several sets of comparison experiments. The validation set and training set are divided by 1:9, the initial learning rate is 0.0004, the training batch size is taken as 10, the maximum number of iterations is 1000, the optimizer is Adaptive moment estimation (Adam), and the learning decay rate is 0.5. During the training process, the corresponding model weight parameters are saved after each iteration to continue training at unexpected training breakpoints. At the

same time, the change of the validation loss value is monitored: the learning rate reduction strategy is triggered when the model performance does not improve in five iterations, and the training is terminated to avoid overfitting when the model performance does not improve in 10 iterations.

In this paper, we use mean Average Precision (mAP) as the evaluation criteria, which is the average of the *precision* on the *precision-recall* curve, and the formula is defined as:

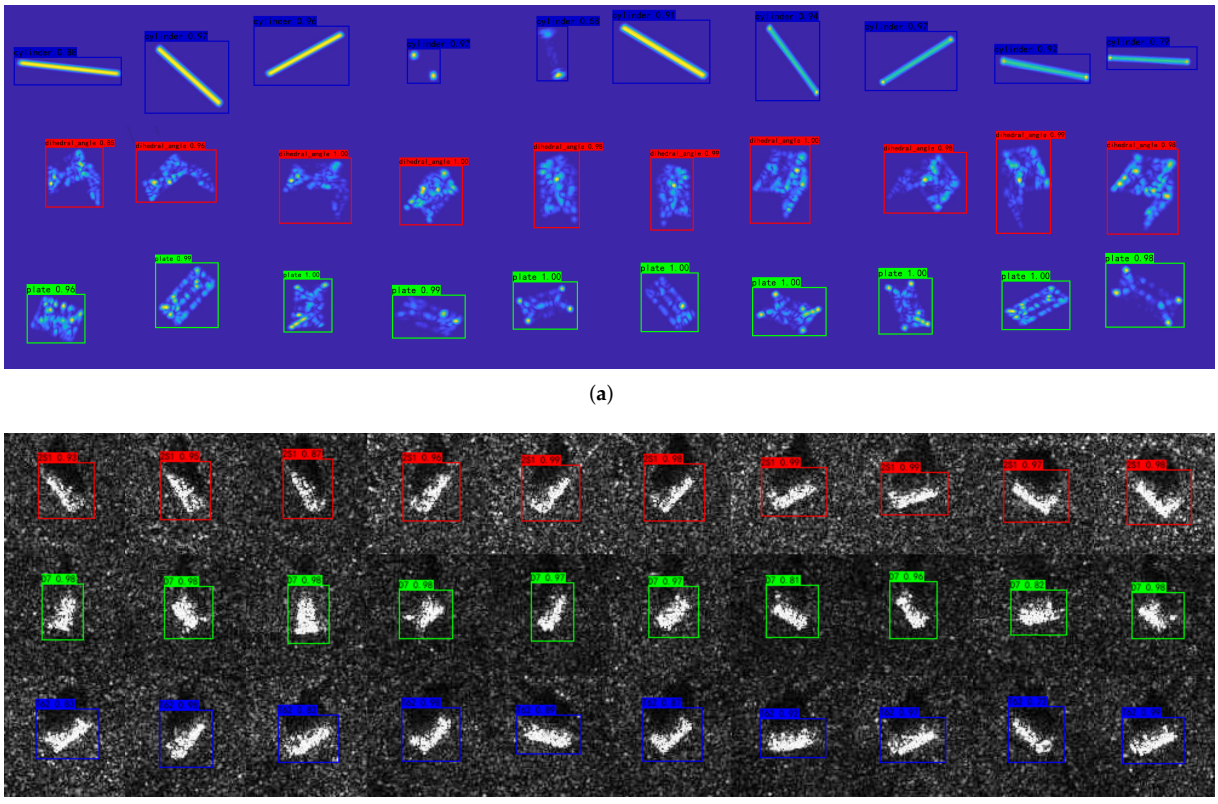
$$\mathrm{AP} = \int_0^1 p(r)dr \tag{5}$$

$$\mathrm{mAP} = \frac{1}{N}\sum_1^N \mathrm{AP} \tag{6}$$

where p is *precision*, r is *recall*, AP is the average precision of one category, and mAP is the mean of average precision of all categories.

3.3. Experimental Results

As Table 3 shows, our method has an average improvement of 3.63% mAP compared with the original SSD. Compared with the Faster R-CNN and YOLOv3 detection models, ours also has the optimal performance, with the highest AP for each category on both datasets. On the GO dataset, ours surpasses in the AP from 0.31% to 6.58% for dihedral angles, surface plates and cylinders, and from 0.05% to 5.69% for 2S1, D7 and T62 on the MSTAR dataset. The experimental results illustrate that the performance of the improved model is more advantageous compared with Faster R-CNN and YOLOv3, which verifies the effectiveness of the improved method. Some of the detection results are given in Figure 7.



(**a**)

(**b**)

Figure 7. Detection results of improved SSD. (**a**) Part of GO dataset test results, (**b**) part of MSTAR dataset test results.

Table 3. Comparison among other detection methods.

AP (%)	GO				MSTAR			
	mAP (%)	Dihedral Angle	Surface Plate	Cylinder	mAP (%)	2S1	D7	t62
Faster R-CNN	96.73	98.54	98.57	93.08	95.98	94.06	99.78	94.10
YOLOv3	96.87	98.89	96.50	95.21	96.39	92.93	99.76	96.47
SSD	94.63	98.09	95.10	91.61	96.01	96.57	98.49	92.97
Ours	**99.50**	**99.96**	**98.88**	**99.66**	**98.40**	**98.62**	**99.83**	**96.75**

3.4. Discussion

In order to verify the performance gains from different components, we conduct separate experiments on the residual connections and small sample augmentation. It can be seen from Figure 8a that, compared to the original feature extraction structure (VGG16), the residual connections have a lead of 2.77% mAP, the GO dataset and 1.86% mAP on the MSTAR dataset, and small sample augmentation has a lead of 1.20% mAP, the GO dataset and 1.74% mAP on the MSTAR dataset. Combining the two components, there are 4.87% and 2.39% mAP enhancement in the two datasets, respectively. On the GO dataset, compared to the small sample augmentation, the residual connections have a stronger contribution to performance gains, while the two methods have the generally equivalent contribution on the MSTAR dataset.

To further validate the reasonableness of aspect ratio improvements as well as saliency map augmentation, Figure 8b discusses the effect of single improvement methods on the backbone-improved model. For cylinder detection, aspect ratios enhance the AP by 4.45%, and saliency map augmentation enhances the AP by 3.71%. On the GO dataset, the combination of the two methods works best for cylinder detection with 8.98% AP and 3.67% mAP. Therefore, the two improved methods complement each other, and the method that we propose, which is the best, considers both optimization of the default box aspect ratios and saliency map augmentation.

Original SSD adopts VGG16 as the backbone, with a total of 13 layers (excluding the Full Connection layer) and a memory size of 56.13MB. We improve the backbone architecture with a residual network, which has a total of 40 layers and a memory size of 26.20MB. From Table 4 and the results of the comparative experiment, it can be seen that even if the residual network deepens the depth of backbone layers, the memory occupied is reduced up to 53.32%, and it has better detection performance.

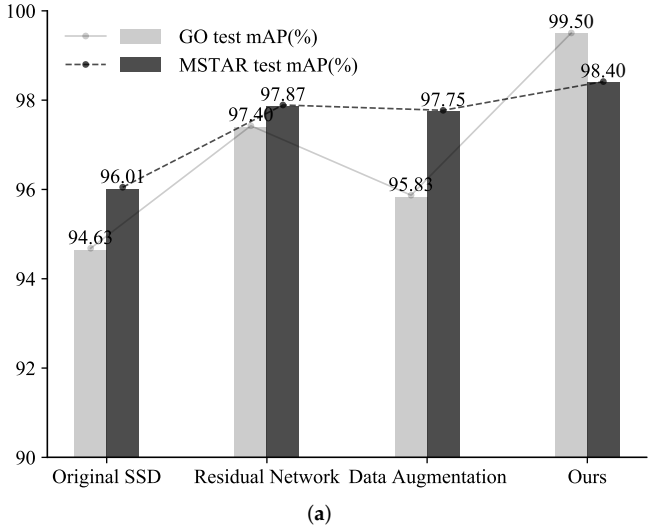
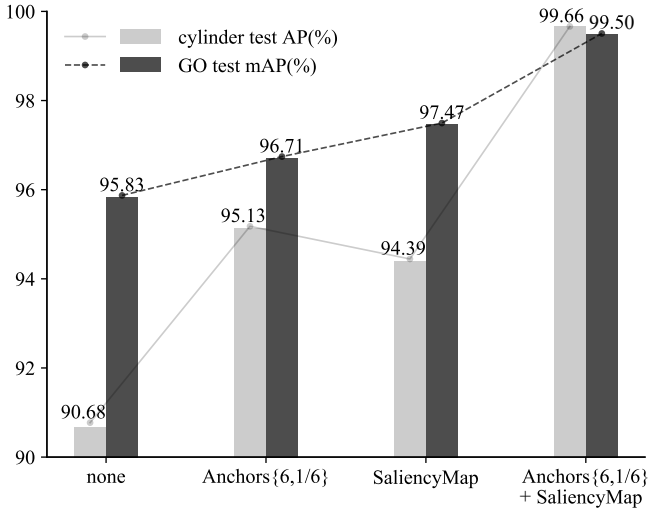


Figure 8. *Cont.*



(**b**)

Figure 8. Effect of different improvement methods on model performance. (**a**) This shows the performance gains from residual connections and small sample augmentation; the final method adopts the two components. (**b**) This shows effectiveness of aspect ratios and saliency map augmentation based on the improved model.

Table 4. Comparison of the memory occupied by VGG16 and residual network.

Backbone	Layers	Memory Size (MB)	Input Size
VGG16	13	56.13	$300 \times 300 \times 3$
residual network	40	26.20	$600 \times 600 \times 3$

4. Conclusions

In this paper, an improved SSD model for SAR target detection is proposed. Differring from the plain feature extraction network, we use a residual network that can extract deeper feature information. To match specific detection targets, we redesign the aspect ratios of default boxes. A small sample augmentation based on the image saliency map theory is proposed to enhance the model generalization ability. The comparison experiments based on the electromagnetic simulation image dataset and MSTAR dataset verify the effectiveness of the improved method, which can achieve better results in SAR target detection.

Author Contributions: Conceptualization, F.Z. and M.X.; methodology, F.Z. and F.H.; validation, F.Z., F.H. and C.G.; formal analysis, F.Z. and F.H.; investigation, Z.D.; data curation, F.H. and C.G.; writing—original draft preparation, F.Z. and F.H.; writing—review and editing, F.Z. and F.H.; supervision, M.X.; project administration, F.Z.; funding acquisition, F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Aeronautical Science Foundation of China under grant number 2019200P4001 and the National Natural Science Foundation of China under grant number 61701156.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Chen, J.; Zhang, J.; Jin, Y.; Yu, H.; Liang, B.; Yang, D.G. Real-Time Processing of Spaceborne SAR Data With Nonlinear Trajectory Based on Variable PRF. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
2. Chen, J.; Xing, M.; Yu, H.; Liang, B.; Peng, J.; Sun, G.C. Motion Compensation/Autofocus in Airborne Synthetic Aperture Radar: A Review. *IEEE Geosci. Remote Sens. Mag.* **2021**, 2–23. [CrossRef]
3. Yang, T.; Li, S.; Xu, O.; Li, W.; Wang, Y. Three dimensional SAR imaging based on vortex electromagnetic waves. *Remote Sens. Lett.* **2018**, *9*, 343–352. [CrossRef]
4. Kuo, J.M.; Chen, K.S. The application of wavelets correlator for ship wake detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1506–1511. [CrossRef]
5. Kang, M.; Baek, J. SAR Image Change Detection via Multiple-Window Processing with Structural Similarity. *Sensors* **2021**, *21*, 6645. [CrossRef]
6. Gao, Y.; Gao, F.; Dong, J.; Wang, S. Change Detection From Synthetic Aperture Radar Images Based on Channel Weighting-Based Deep Cascade Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 4517–4529. [CrossRef]
7. Joshi, S.K.; Baumgartner, S.V.; da Silva, A.B.C.; Krieger, G. Range-Doppler Based CFAR Ship Detection with Automatic Training Data Selection. *Remote Sens.* **2019**, *11*, 1270. [CrossRef]
8. Huang, Y.; Liu, F. Detecting Cars in VHR SAR Images via Semantic CFAR Algorithm. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 801–805. [CrossRef]
9. Fan, J.; Tomas, A. Target Reconstruction Based on Attributed Scattering Centers with Application to Robust SAR ATR. *Remote Sens.* **2018**, *10*, 655. [CrossRef]
10. Zhu, J.; Qiu, X.; Pan, Z.; Zhang, Y.; Lei, B. Projection Shape Template-Based Ship Target Recognition in TerraSAR-X Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 222–226. [CrossRef]
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014. [CrossRef]
12. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
14. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
15. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
17. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
18. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
19. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
20. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. Automatic Ship Detection Based on RetinaNet Using Multi-Resolution Gaofen-3 Imagery. *Remote Sens.* **2019**, *11*, 531. [CrossRef]
21. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional Neural Network with Data Augmentation for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 364–368. [CrossRef]
22. Gao, F.; Shi, W.; Wang, J.; Yang, E.; Zhou, H. Enhanced Feature Extraction for Ship Detection from Multi-Resolution and Multi-Scene Synthetic Aperture Radar (SAR) Images. *Remote Sens.* **2019**, *11*, 2694. [CrossRef]
23. Lin, Z.; Ji, K.; Kang, M.; Leng, X.; Zou, H. Deep Convolutional Highway Unit Network for SAR Target Classification with Limited Labeled Training Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1091–1095. [CrossRef]
24. He, Z.; Xiao, H.; Tian, Z. Multi-View Tensor Sparse Representation Model for SAR Target Recognition. *IEEE Access* **2019**, *7*, 48256–48265. [CrossRef]
25. Wang, J.; Zheng, T.; Lei, P.; Bai, X. A Hierarchical Convolution Neural Network (CNN)-Based Ship Target Detection Method in Spaceborne SAR Imagery. *Remote Sens.* **2019**, *11*, 620. [CrossRef]
26. Wang, Z.; Du, L.; Mao, J.; Liu, B.; Yang, D. SAR Target Detection Based on SSD With Data Augmentation and Transfer Learning. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 150–154. [CrossRef]
27. Zhang, X.; Liu, G.; Zhang, C.; Atkinson, P.M.; Tan, X.; Jian, X.; Zhou, X.; Li, Y. Two-Phase Object-Based Deep Learning for Multi-Temporal SAR Image Change Detection. *Remote Sens.* **2020**, *12*, 548. [CrossRef]
28. Chen, H.; Shi, Z. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]