sustainability

MDPI

Article

Resource-Based Port Material Yard Detection with SPPA-Net

Xiaoyong Zhang [1], Rui Xu [1], Kaixuan Lu [2], Zhihang Hao [1], Zhengchao Chen [2] and Mingyong Cai [3,4,*]

[1] Beijing Key Laboratory of High Dynamic Navigation, Beijing Information Science and Technology University, Beijing 100101, China
[2] Airborne Remote Sensing Center, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[3] State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China
[4] Satellite Application Center for Ecology and Environment, MEE, Beijing 100094, China
* Correspondence: caimingyong@e-mail.secmep.cn

Abstract: Since the material yard is a crucial place for storing coal, ore, and other raw materials, accurate access to its location is of great significance to the construction of resource-based ports, environmental supervision, and investment and operating costs. Its extraction is difficult owing to its small size, variable shape, and dense distribution. In this paper, the SPPA-Net target detection network was proposed to extract the material yard. Firstly, a Dual-Channel-Spatial-Mix Block (DCSM-Block) was designed based on the Faster R-CNN framework to enhance the feature extraction ability of the location and spatial information of the material yard. Secondly, the Feature Pyramid Network (FPN) was introduced to improve the detection of material yards with different scales. Thirdly, a spatial pyramid pooling self-attention module (SPP-SA) was established to increase the global semantic information between material yards and curtail false detection and missed detection. Finally, the domestic GF-2 satellite data was adopted to conduct extraction experiments on the material yard of the port. The results demonstrated that the detection accuracy of the material yard reached 88.7% when the recall rate was 90.1%. Therefore, this study provided a new method for the supervision and environmental supervision of resource-based port material yards.

Keywords: material yard detection; deep learning; attention mechanism

Citation: Zhang, X.; Xu, R.; Lu, K.; Hao, Z.; Chen, Z.; Cai, M. Resource-Based Port Material Yard Detection with SPPA-Net. *Sustainability* **2022**, *14*, 16413. https://doi.org/10.3390/su142416413

Academic Editor: Ripon Kumar Chakrabortty

Received: 28 October 2022
Accepted: 3 December 2022
Published: 8 December 2022

1. Introduction

Ports, as a hub of maritime logistics, undertake the function of the temporary storage of coal, ore, grain, and other raw materials. After entering ports, these raw materials will be transferred to the material yard for temporary stacking and then transported to other destinations by train or ship [1]. In 2018, the Ministry of Transport announced The Action Plan for Further Promoting Green Port Construction (2018–2022) [2], which was proposed to integrate the concept of green development into the construction of ports. However, the ports themselves and the surrounding environment are seriously influenced due to the considerable amount of dust generated in the daily construction process of open-air material yards of ports. Given the above reasons, the detection of port material yards can not only assist in the planning of their development but also facilitate the timely and effective supervision of open material yards in ports by environmental protection departments in various regions.

Compared with the two port targets (ships and oil storage tanks), material yards are characterized by small targets, dense distribution, and variable scale. The shape of ships and storage tanks is fixed, while the shape and size of material yards usually change with the stacking process of the workers. Some have an excessively large scale, while some have an excessively small scale. Some are rectangular, some are elliptical, and some are irregular. These factors lead to the difficulty in detecting material yard targets.

In recent years, deep learning, as a sample-driven data analysis method, has been extensively used in the field of remote sensing. The deep convolutional neural network (CNN) considerably improves the recognition accuracy of the image. It does not require tedious manual design while being able to autonomously perceive the feature information in the image and present better universality and expansibility. There are two main types of object detection methods based on deep learning: (1) single-stage target detection algorithms, including SSD [3], YOLO series [4–7], and RetinaNet [8]; (2) two-stage target detection algorithms, such as R-CNN [9], Fast R-CNN [10], and Faster R-CNN [11]. The composition of the two-stage target detection algorithms mainly consists of the feature extraction network and the region recommendation network in the first stage, as well as the classifier and regressor in the second stage. Feature extraction networks are employed to extract the feature information of the target. The commonly used feature extraction networks comprise VGGNet [12], GoogleNet [13], and ResNet [14]. ResNet tackles the phenomenon of the exploding gradient or vanishing gradient of the network with the increase in the number of layers to a certain extent [14]. The region proposal network is the core of two-stage target detection algorithms, and this is the fundamental difference from one-stage algorithms. Its role is to generate a series of candidate boxes that may contain targets and roughly screen the original image. The effect of dense detection can be achieved by laying a large number of candidate boxes on the original image. There is no region proposal network in the one-stage algorithm, but feature extraction is directly performed on the image to predict the location and classification of the target object. Therefore, the accuracy of the two-stage algorithm is higher than that of the one-stage algorithm. An attention mechanism [15] is needed to effectively extract image features. It is the embodiment of selective attention in the field of computer vision. It assigns different weights to the feature maps, and the positions with more weights represent more attention. Among the current attention mechanisms, channel attention [16], spatial attention [17–19], channel-attention attention [20,21] and self-attention [22] have been widely used.

Several target detection algorithms based on deep learning have been proposed to handle small, densely distributed, and multi-scale remote sensing image targets. For example, Lu et al. [23] utilized the hybrid attention mechanism of spatial attention and channel attention mechanism in parallel to effectively suppress the background noise of the image and strengthen the feature extraction ability. The mAP of small ground targets such as vehicles and ships reached 52.6%. Hua et al. [24], Ying et al. [25], and Zhu et al. [26] added self-attention mechanisms to different network structures to improve the extraction accuracy of small densely distributed targets. Huang Zhipeng et al. [27] enhanced the Faster R-CNN by sending the feature maps generated in different stages of the feature extraction network into the RPN to obtain the feature information of targets with different sizes. Its accuracy was boosted by 5% compared with the original Faster R-CNN. Lin Zhijie et al. [28] further improved the Faster R-CNN based on Huang Zhipeng et al. and employed the feature maps generated by the last three stages of the feature extraction network to construct a feature pyramid, contributing to reinforcing the detection ability of multi-scale targets. Its mAP on the PASCAL VOC 2007 and 2008 datasets reached 74.8%. Li et al. [29] improved the feature pyramid and proposed a saliency-based pyramid combining the feature pyramid and the saliency algorithm, which augmented the ability to reduce background noise. The mAP of the aerial image data set reached 72.96%. Zhong et al. [30] added the structure of the feature pyramid and self-attention mechanism to YOLOv3 to strengthen the detection ability of multi-scale and densely distributed targets, and the mAP reached 87.41% on the UCAS-AOD dataset.

Although the improved methods mentioned above have improved the detection accuracy of different targets, material fields with different shapes and complex background information will cause the problem of missed detection and false detection under the conditions of multi-scale and densely distributed remote sensing images. Therefore, an SPPA-Net target detection algorithm was proposed in this paper for the detection of port material yards based on the domestic GF-2 satellite remote sensing image. This method

adopts the Faster R-CNN and the ResNet-50 as the basic framework and the feature extraction network, respectively. Firstly, the dual hybrid attention module was proposed in this paper to increase the effectiveness of the channel and spatial information extraction in the feature extraction stage. Secondly, the feature pyramid structure was introduced, and then the spatial pyramid pooling self-attention module proposed in this paper was integrated to enrich the semantic information of the feature map for subsequent network detection. Finally, the target detection data set of self-built port material yards was adopted to train the algorithm. The experimental results suggested that the proposed method effectively extracted the material field target, with a recall rate and accuracy rate of 90.1% and 88.7%, respectively.

The main contributions of this paper are as follows:

(1) A deep learning target detection algorithm was constructed for port stockyard targets, and the algorithm was used to verify with self-built port material yard datasets;
(2) The Dual-Channel-Spatial-Mix Block was proposed, to improve the feature extraction ability of densely arranged and multi-scale stockyard targets;
(3) The spatial pyramid pooling attention module was designed to globally model the features of each position in the feature map for obtaining more abstract global features.

2. Principles and Methods

2.1. Sppa-Net

The frame foundation of the SPPA-Net is the Faster R-CNN, as shown in Figure 1, which consists of four parts: feature extraction network, Region Proposal Networks (RPN), ROI Pooling layer, and parallel classifiers and regressors. The algorithm flow is detailed, as follows. First, the image is input into the feature extraction network to obtain the feature map; second, the feature map is input into the RPN to acquire a candidate box that may be the target; third, the matrix of the feature map of the image area where the candidate box is located is scaled to a 7 × 7 feature map through the ROI Pooling layer; finally, the scaled feature map matrix is input into the classifier and regressor to generate the predicted results.

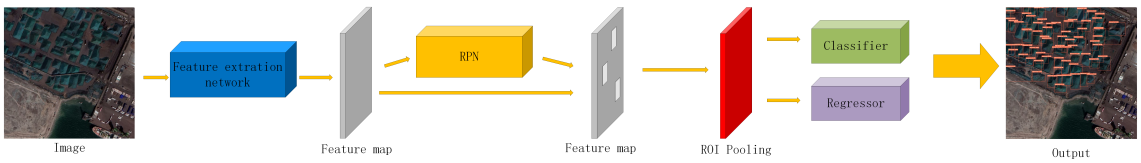

Figure 1. The overall framework of the SPPA-Net.

In this paper, the ResNet-50 was selected as the feature extraction network, which consists of five stages. Specifically, stage 1 is composed of convolution and maximum pooling, and the remaining stages are stacked by residual structures. The region proposal network consists of a fully convolutional network. The classifier and regressor comprise a fully connected layer.

In this paper, the feature extraction network was improved from the following aspects to enhance the extraction ability of densely distributed, multi-scale, and small targets. (1) The dual hybrid attention mechanism module proposed in this paper was added after each residual structure of the ResNet-50; (2) the feature pyramid was constructed using the feature map generated by the ResNet-50 in each stage, and a spatial pooling self-attention mechanism was added after the feature map output by the feature pyramid. The detailed structure is illustrated in Figure 2.

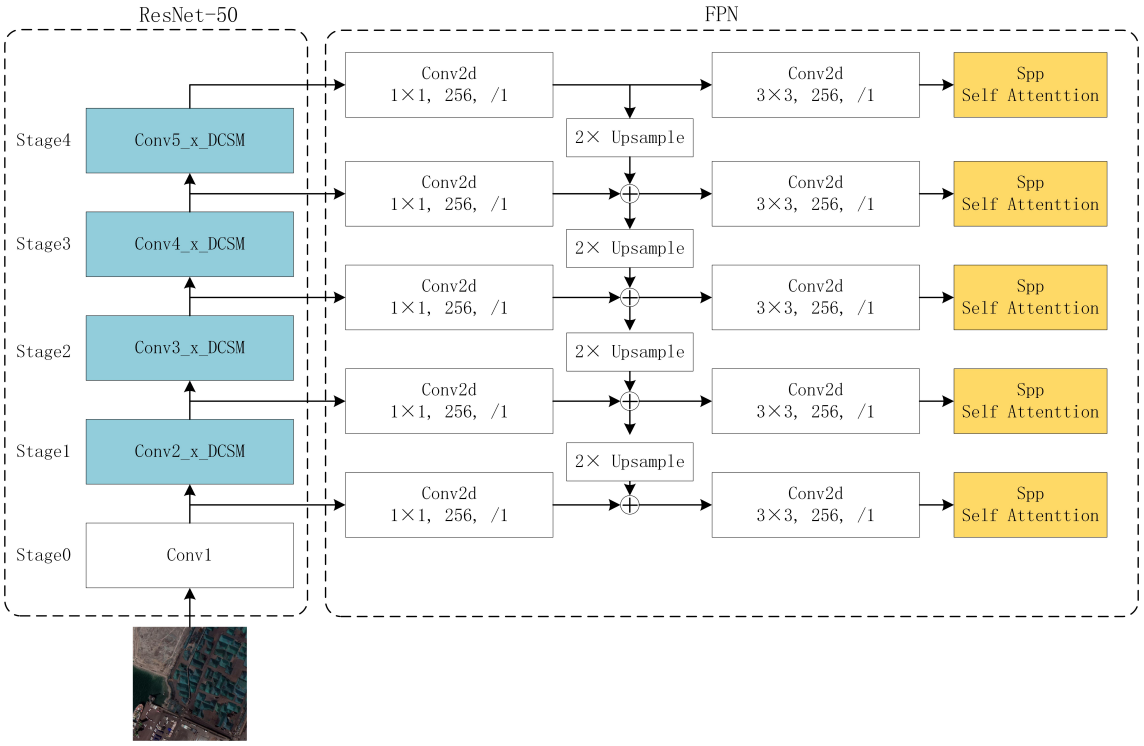

Figure 2. Improved feature extraction network.

2.2. Improved Feature Extraction Network

2.2.1. Double-Mixed Attention Mechanism

The shape of various material yards is noticeably different. Moreover, the number of material yards, the types of materials stacked, and the arrangement of materials significantly vary in different scenarios. Hence, a dual-mixed attention mechanism was proposed in this paper to improve the network's ability to extract features of material yards.

Mixed attention is composed of channel attention and spatial attention. Generally, there are two ways of combination: series and parallel. On this basis, a dual-mixed attention module was constructed. The structure of the DCSM-Block is exhibited in Figure 3. It consists of two horizontal attention modules and two spatial attention modules, which are combined in series and parallel simultaneously.

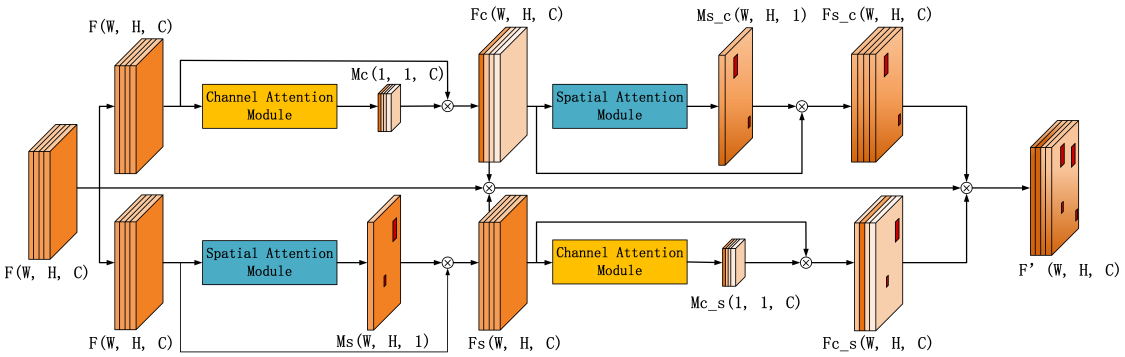

Figure 3. Structure of Dual-Channel-Spatial-Mix Block.

After entering the DCSM-Block, the feature map F with a size of W×H×C entered the upper and lower lines, as presented in Figure 3. On the upper line, the feature map *F* will first enter the channel attention module to perform feature extraction on the channel dimension information, and finally obtain the feature map *Fc* with different channel weight information. Then, *Fc* will be sent to the spatial attention module. The spatial

attention module will perform further feature extraction on the feature map *Fc* with channel information, so as to obtain the position feature information of the target to be detected, and finally generate a feature map *Fs_c* with both the target position information and channel information. The specific process is as follows:

$$M_c = \sigma(MLP(MaxPool(F)) + MLP(AvgPool(F))), \quad (1)$$

$$F_c = F \otimes M_c, \quad (2)$$

$$M_{s\_c} = \sigma(f^{7\times7}(AvgPool(F_c); MaxPool(F_c)) = \sigma(f^{7\times7}(F_{c\_pool})), \quad (3)$$

$$F_{s\_c} = F_c \otimes M_{s\_c}, \quad (4)$$

where M_c represents the channel attention map; $\otimes$ represents the dot multiply operator; σ represents the Sigmiod function; MLP represents the multi-layer perceptron; MaxPool represents the global maximum pooling; AvgPool represents the global average pooling; M_{s_c} represents the spatial attention map with channel information; $f^{7\times7}$ represents 7×7 convolution layer.

On the lower line, the spatial attention module and the channel attention module performed the same operation on the feature map F as on the upper line. Firstly, the feature map F generated the feature map Fs with spatial information through the spatial attention module. Secondly, the feature map Fs entered the channel attention module to generate a feature map F_{c_s} with channel and spatial information. Finally, the feature maps F, F_c, F_s, F_{s_c}, and F_{c_s} were multiplied to obtain the feature map F′, so as to further strengthen the feature information of the space and channel.

$$F' = F \otimes F_c \otimes F_s \otimes F_{s\_c} \otimes F_{c\_s}, \quad (5)$$

In this paper, DCSM-Block was added after each residual structure of ResNet-50 to increase the content of critical parts of the feature map and curtail the features of useless information.

2.2.2. Spatial Pyramid Pooling Self-Attention Module

In ResNet-50, a feature map with a gradually decreasing size and a gradually increasing number of channels was generated. The feature map of the early stage contained the feature information of the small target. Nevertheless, the feature information of the small target in the feature map generated in the later stage was ignored, while the feature information of the large target was retained. Therefore, a feature pyramid structure based on the feature maps generated by ResNet-50 at each stage was constructed for the semantic information of small targets to strengthen the extraction ability of small-size material yards. Although the feature pyramid structure can significantly improve the detection effect of small targets in material yards, it does not wrestle from the non-target false detection of the network. Therefore, in this paper, spatial pyramid pooling was introduced into the self-attention mechanism, and this improved self-attention mechanism was added after the feature map output by the feature pyramid to make up for the shortcomings of the local perception of convolutional neural networks.

The core content of the self-attention module, as displayed in Figure 4a, is to calculate the relationship between the pixels in the feature map and to achieve global context modeling. Different from the spatial attention mechanism, the self-attention mechanism not only assigns weights to each pixel in the feature map using a single-layer convolution structure but also maps the original feature image into three vector branches (Query, Key, and Value). The self-attention mechanism is different from the one-dimensional vector in natural language processing, though it can effectively model the relationship between each pixel in the feature map. Images are the basic input in computer vision, and the generated one-dimensional vector is too long in the process of vectorization, resulting in a serious waste of computing resources. With the feature map size of 200×200 input as an example,

the red box in Figure 4a will generate dimensions of (40,000, 40,000), which will take up a lot of computing resources. This not only puts forward higher requirements for hardware but also makes it impossible to train in batches ascribed to the large occupation of memory, reducing the speed of model convergence.

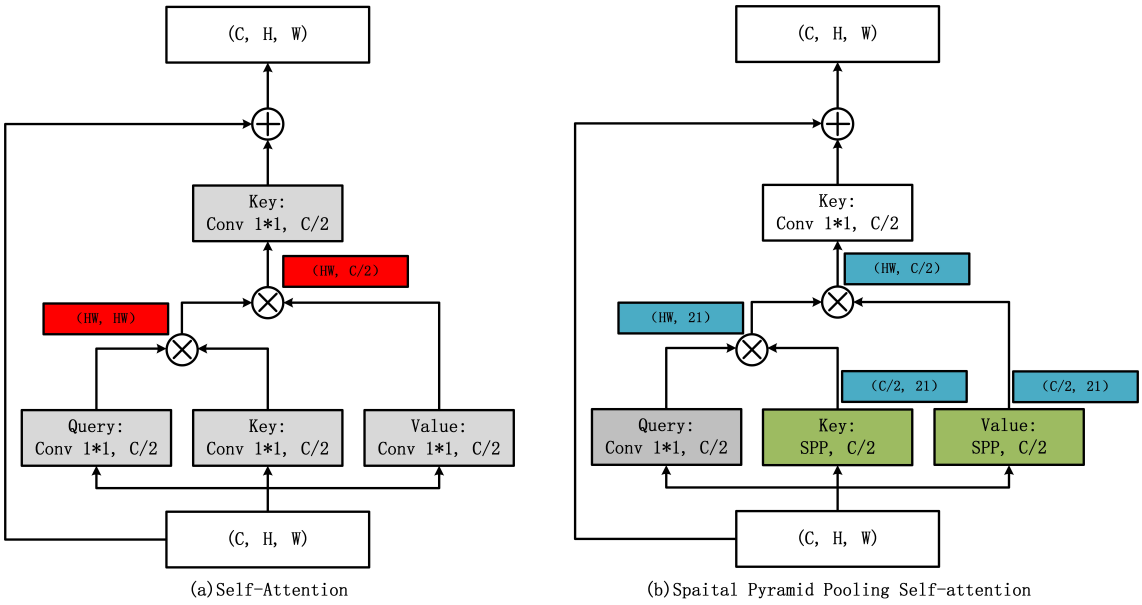

Figure 4. Comparison between self-attention and spatial pyramid pooling self-attention module: (**a**) Self-attention module. (**b**) Spatial pyramid pooling self-attention module. (SPP represents Spatial pyramid pooling; the red rectangular box represents the change in the dimension of the feature matrix in the self-attention module; the blue rectangle represents the dimension change of the feature matrix of the spatial pyramid pooling self-attention module).

Aiming at the above problems of self-attention, the dimension was lowered in this paper without losing feature information by improving the two branches of Key and Value to reduce the amount of computation and memory usage, so as to grapple with the above complications of the self-attention module. As suggested in Figure 4b, the 1 × 1 convolutional layers of the Key and Value branches were replaced by the spatial pyramid pooling layers, and then the feature maps were extracted at different resolutions through pooling windows of different sizes to form a one-dimensional feature vector. Compared with the 1 × 1 convolution in the original self-attention module, it is easier for the spatial pyramid pooling network to extract the global semantic information of the feature map, and the feature dimension obtained was much smaller than the result after 1 × 1 convolution processing. Compared with the self-attention module, the spatial pyramid pooling self-attention module has significantly reduced the amount of computation. From the experimental results, the precision rate and recall rate have been significantly improved.

3. Results and Discussion

3.1. Dataset

In this paper, GF-2 satellite remote sensing images of Tianjin Port and Tangshan Port were used to prepare the data set. The specific workflow is displayed in Figure 5. Firstly, ArcGIS software was employed to remove the non-port area and thus obtain the overall remote sensing images of the two ports. Secondly, Python was adopted to write a script to cut the original remote sensing image into slices of the same size, and the size of the slices was 1024 × 1024. Finally, the number of slices in the Tangshan port area and the Tianjin port area was 1872 and 9828, respectively.

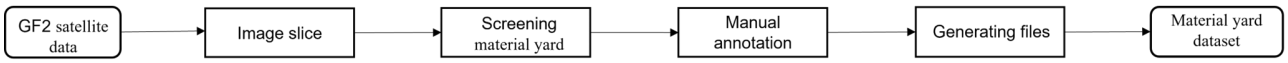

Figure 5. Material yard sample preparation.

The slices containing material field targets were selected from the slices of the Tianjin Port and Tangshan Port with manual interpretation as the data set. Finally, 1362 images were obtained, and the number of material fields was 10,191. Deep learning should use labeled training data to train the algorithm. In this paper, LabelMe software was utilized to mark the sample pictures, and polygons were adopted to mark the material yard to obtain accurate edge information. Each sample image corresponded to a tag file. Finally, the tag files were summarized using Python to generate training files and verification files.

Through the above steps, 1362 material yard data sets with a size of 1024 × 1024 were obtained, among which 1225 and 137 were used as the training set and the test set, respectively, at a ratio of 9:1.

3.2. Environment Configuration and Training Methods

The environment hardware device used in this training is an NVIDIA graphics card, with a model of Titan XP and a memory size of 12196MiB. The software involves CUDA (version 11.2), Python (version 3.7), and PyTorch (deep learning framework 1.10). Random data enhancement was firstly performed on the read training samples to expand the training samples. The main data enhancement methods include random cropping, flipping, brightness transformation, and contrast transformation. Finally, the data enhanced was made into variables in PyTorch for gradient calculation of backpropagation. The hyperparameters of the network are comprised of the optimizer, Lr config, and epoch. The details are shown in Table 1.

Table 1. Hyper-parameters setting.

The Hyperparameters	Parameters Setting	
optimizer	Type	SGD
	Learning base	0.02
	Momentum	0.9
	Weight decay	0.0001
Lr config	Policy	Step
	Warmup	Linear
	Warmup iters	500
	Warmup ratio	0.001
	Step	8
epoch	120,000	

The metrics including *Precision*(*P*), *Recall* (*R*), and *mAP* were adopted to evaluate the network performance constructed in this paper, expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$mAP = \int_0^1 P(R)dR \tag{8}$$

where TP (True Positives) represents the number of actual material yard targets correctly identified as material yards, FP (False Positives) indicates the number of the actual backgrounds but are mistakenly identified as material yards, and FN (False Negatives) denotes the number of actual material yard targets but are mistakenly classified as the background. Precision reflects the correct proportion of all material yard targets predicted by the model. Recall implies what proportion of all material yard targets is predicted by the model.

In addition, mAP (mean Average Precision) signifies the average accuracy rate, which comprehensively evaluates the accuracy and recall rates of the model. The calculation method is to calculate the accuracy and recall rates under different IOU thresholds, and then draw a curve with abscissa and ordinate. Finally, the area enclosed by the curve and abscissa and ordinate is calculated, where the IOU threshold is from 0.5 to 0.95, increasing every 0.05.

3.3. Ablation Experimental Results and Analysis

In this paper, ablation experiments were conducted to verify the effectiveness of the mixed attention mechanism and the FPN structure with the improved self-attention mechanism. Firstly, the Faster R-CNN network was used as the baseline network, with the mixed attention mechanism DCSM-Block and the FPN with SPP-SA module (SPP-SA FPN) being added, respectively. As shown in Table 2, the results revealed that the mAP score and recall rate were significantly increased by 3.2% and 5.2%, respectively, with the superposition and improvement of the two modules.

Table 2. Precision, Recall and mAP after superposition of DCSM and SPP-SA FPN modules.

DCSM	SPP-SA FPN	Precision	Recall	mAP
-	-	0.860	0.824	0.881
✓	-	**0.888**	0.845	0.897
✓	✓	0.887	**0.901**	**0.913**

Note: - represents that the module was not added to the experiment; ✓represents the module was added to the experiment.

In Figure 6(1), the original Faster R-CNN falsely detected targets similar in color to the material yard; in Figure 6(2)–(5), the Faster R-CNN had a significant missed detection when the material yard was arranged too densely and the shape was changed. After the Faster R-CNN of the DCSM-Block was added, the false detection and missed detection rates of the material yard were significantly reduced, and the accuracy and recall rates were improved by 3.2% and 2.1%, respectively. This was in that the DCSM-Block module enhanced the ability of the backbone network to extract the characteristics of the material field and focused the attention of the algorithm on faceted objects of the material yard.

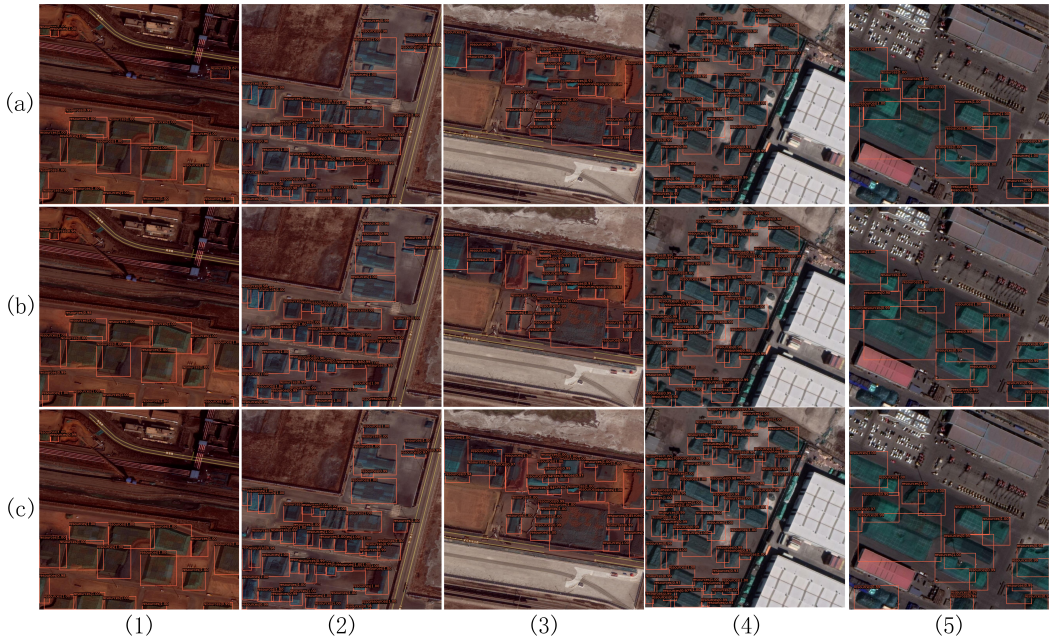

Figure 6. Results after superposition of DCSM and SPP-SA FPN modules: (**a**) Faster R-CNN. (**b**) Faster R-CNN + DCSM Block. (**c**) Faster R-CNN + DCSM module + SPP-SA FPN module.

After the addition of SPP-SA FPN, the accuracy slightly decreased compared with the DCSM-Block, the recall rate was significantly improved by 5.6%, and the detection ability of small-area material yards was further improved, as illustrated in Figure 6(2)–(5). The reason for this phenomenon is that the self-attention mechanism boosted the network's perception of global information and reinforced the classification ability of the algorithm while weakening the positioning ability of the algorithm, leading to a decrease in the accuracy of the model and an increase in the recall rate.

3.4. Comparative Experimental Results and Analysis

SSD300, SSD512, YOLOv7, RetinaNet, Faster R-CNN, and the proposed method were compared to verify the effectiveness of the algorithm. The experimental results are listed in Table 3. The proposed algorithm achieved the highest recall rate and mAP (0.901 and 0.913), respectively. SSD512 reached the highest precision and parameter (0.907 and 46.04 M).

Table 3. Comparison of Precision, Recall, mAP and Parameters of different algorithms.

Algorithms	Precision	Recall	mAP	Parameters
SSD300	0.906	0.701	0.857	34.31 M
SSD512	**0.907**	0.785	0.906	**46.04 M**
YOLOv7	0.890	0.816	0.905	37.20 M
RetinaNet	0.782	0.707	0.740	37.74 M
Faster R-CNN	0.860	0.824	0.881	41.12 M
Ours	0.887	**0.901**	**0.913**	41.53 M

The specific test results are demonstrated in Figure 7, from which the following findings can be obtained.

(1) These six methods can extract large and rectangular material fields. However, there are some material fields with irregular shapes in Figure 7(3). Except for the method in this paper, a certain degree of missed detection occurred, among which RetinaNet was the most significant. This suggested that the DCSM-Block in this paper effectively extracted the feature information of the material field, reduced the useless feature information, and thus enhanced the network's ability to extract the characteristics of the material field.

(2) Concerning small and densely-distributed material yards, as illustrated in Figure 7(2) and (4), the most severe missed detection of RetinaNet and SSD300 also missed a large area of the material yard. In SSD512 and YOLOv7, the phenomenon of missed detection of large-area stockyards occurred less, while small-area stockyards when the distance between stockyards was close was not detected. Additionally, YOLOv7 caused the false detection of targets, whose color was similar to that of the material field. The method in this paper demonstrated superiority in this respect. As the FPN structure was constructed in this paper and the SPP-SA module was added after the feature map, it improved the global perception ability, effectively enhanced the extraction ability of multi-scale and small-area material fields, and compensated for the local perception of the convolutional neural network.

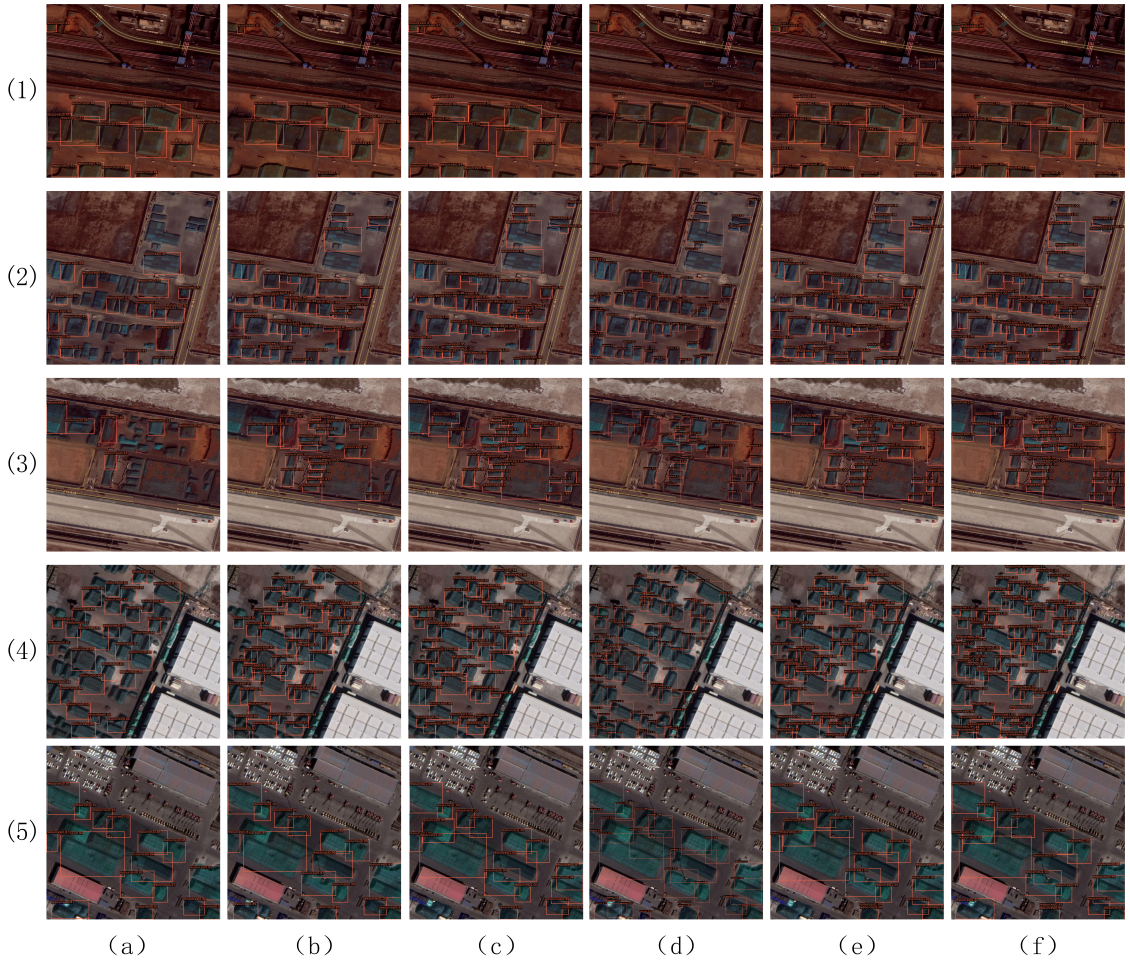

Figure 7. Comparison of different algorithms: (**a**) RetinaNet. (**b**) SSD300. (**c**) SSD512. (**d**) YOLOv7. (**e**) Faster R-CNN. (**f**) Ours.

4. Conclusions

In this paper, the SPPA-Net algorithm was proposed to detect material yard targets. Based on the original Faster R-CNN, ResNet-50 was selected as the feature extraction network, and a dual mixed attention module was embedded to enhance the extraction of the material field features. Subsequently, the feature pyramid was constructed using the feature maps generated by the ResNet-50 at each stage. The spatial pyramid pooling self-attention module was embedded to globally model the features of each position in the feature map, so as to compensate for the limitations of the local perception of the convolutional neural networks and expand the universality of the network. Compared with the original attention mechanism, the computational complexity was reduced by 19 times. Finally, a material field data set was established with the GF-2 satellite. The experiment revealed that the proposed method enabled fast and efficient extraction of the material fields within ports with high accuracy. Compared with other methods, this paper effectively improved the extraction of densely distributed and variable-scale stockyard targets, while curtailing the probability of false detection and missed detection. The recall rate reached 90.1%, and the accuracy rate reached 88.7%. In this paper, the single objective of the material yard was only employed to verify the effectiveness of the proposed algorithm. In future research, target detection will be performed on important targets in other ports, such as ships, containers, oil storage tanks and wharves, so as to further demonstrate the performance of the method proposed in this paper.

Author Contributions: Conceptualization, R.X.; Methodology, R.X. and Z.H.; Data curation, K.L.; Investigation, R.X.; Resources, X.Z.; Visualization, R.X.; Writing—original draft, R.X. and X.Z.;

Review and editing, X.Z., Z.C. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Key Research and Development Program of China (Grant No. 2021YFB3901202).

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. Zhang, Q.; Wang, S.; Zhen, L. Yard truck retrofitting and deployment for hazardous material transportation in green ports. *Ann. Oper. Res.* **2022**, *in press*. [CrossRef]
2. Xi, Y. The Action Plan for Further Promoting Green Port Construction. *China Logist. Purch.* **2018**, *2*, 33–34. 10.16079/j.cnki.issn1671-6663.2018.08.005. [CrossRef]
3. 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, 11–14 October 2016; pp. 21–37.
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
5. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
6. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
7. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
8. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
10. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
12. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
13. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
14. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
16. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [CrossRef]
17. Xu, Y.; Zhang, Y.; Yu, C.; Ji, C.; Yue, T.; Li, H. Residual Spatial Attention Kernel Generation Network for Hyperspectral Image Classification with Small Sample Size. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3175494. [CrossRef]
18. Praveen, B.; Menon, V. Dual-Branch-AttentionNet: A Novel Deep-Learning-Based Spatial-Spectral Attention Methodology for Hyperspectral Data Analysis. *Remote Sens.* **2022**, *14*, 3644. [CrossRef]
19. Peñaloza, B.; Ogmen, H. Effects of spatial attention on spatial and temporal acuity: A computational account. *Attention, Percept. Psychophys.* **2022**, *84*, 1886–1900. [CrossRef] [PubMed]
20. Song, C.H.; Han, H.J.; Avrithis, Y. All the attention you need: Global-local, spatial-channel attention for image retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 2754–2763.
21. Liu, T.; Luo, R.; Xu, L.; Feng, D.; Cao, L.; Liu, S.; Guo, J. Spatial Channel Attention for Deep Convolutional Neural Networks. *Mathematics* **2022**, *10*, 1750. [CrossRef]
22. Cao, F.; Lu, X. Self-attention technology in image segmentation. In Proceedings of the International Conference on Intelligent Traffic Systems and Smart City (ITSSC 2021), Nanjing, China, 28–30 October 2022; Volume 12165, pp. 271–276.

23. Lu, X.; Ji, J.; Xing, Z.; Miao, Q. Attention and feature fusion SSD for remote sensing object detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–9. [CrossRef]
24. Hua, X.; Wang, X.; Rui, T.; Zhang, H.; Wang, D. A fast self-attention cascaded network for object detection in large scene remote sensing images. *Appl. Soft Comput.* **2020**, *94*, 106495. [CrossRef]
25. Ying, X.; Wang, Q.; Li, X.; Yu, M.; Jiang, H.; Gao, J.; Liu, Z.; Yu, R. Multi-attention object detection model in remote sensing images based on multi-scale. *IEEE Access* **2019**, *7*, 94508–94519. [CrossRef]
26. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2778–2788.
27. Huang, J.; Shi, Y.; Gao, Y. Multi-Scale Faster-RCNN Algorithm for Small Object Detection. *Comput. Res. Dev.* **2019**, *56*, 319–327.
28. Lin, Z.; Luo, Z.; Zhao, L.; Lu, D. Multi-scale convolution target detection algorithm with feature pyramid. *J. Zhejiang Univ.* **2019**, *53*, 533–540.
29. Li, C.; Luo, B.; Hong, H.; Su, X.; Wang, Y.; Liu, J.; Wang, C.; Zhang, J.; Wei, L. Object detection based on global-local saliency constraint in aerial images. *Remote Sens.* **2020**, *12*, 1435. [CrossRef]
30. Li, Z.; Wang, H.; Zhong, H.; Dai, Y. Self-attention module and FPN-based remote sensing image target detection. *Arab. J. Geosci.* **2021**, *14*, 1–18. [CrossRef]