PAG-YOLO: A Portable Attention-Guided YOLO Network for Small Ship Detection

Jianming Hu ¹ , Xiyang Zhi ^{1,*} , Tianjun Shi ¹, Wei Zhang ¹, Yang Cui ² and Shenggang Zhao ²

¹ Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China; hujianming@stu.hit.edu.cn (J.H.); 20S021010@stu.hit.edu.cn (T.S.); wzhang@hit.edu.cn (W.Z.)

² Innovation Academy for Microsatellites for Chinese Academy of Sciences, Shanghai 201220, China; cui992579@163.com or 17B921011@stu.hit.edu.cn (Y.C.); sg-zhao@163.com or 18B921010@stu.hit.edu.cn (S.Z.)

* Correspondence: zhixiyang@hit.edu.cn; Tel.: +86-0451-86414883

Abstract: The YOLO network has been extensively employed in the field of ship detection in optical images. However, the YOLO model rarely considers the global and local relationships in the input image, which limits the final target prediction performance to a certain extent, especially for small ship targets. To address this problem, we propose a novel small ship detection method, which improves the detection accuracy compared with the YOLO-based network architecture and does not increase the amount of computation significantly. Specifically, attention mechanisms in spatial and channel dimensions are proposed to adaptively assign the importance of features in different scales. Moreover, in order to improve the training efficiency and detection accuracy, a new loss function is employed to constrain the detection step, which enables the detector to learn the shape of the ship target more efficiently. The experimental results on a public and high-quality ship dataset indicate that our method realizes state-of-the-art performance in comparison with several widely used advanced approaches.

Citation: Hu, J.; Zhi, X.; Shi, T.; Zhang, W.; Cui, Y.; Zhao, S. PAG-YOLO: A Portable Attention-Guided YOLO Network for Small Ship Detection. *Remote Sens.* **2021**, *13*, 3059. <https://doi.org/10.3390/rs13163059>

Academic Editor: Pedro Melo-Pinto

Received: 23 June 2021

Accepted: 30 July 2021

Published: 4 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: YOLO network; attention mechanism; loss function; small ship detection; remote sensing

1. Introduction

Ship detection is of great significance for marine automatic fishery management, port rescue, marine traffic maintenance, and other applications [1–4]. The accuracy of detection is linked to the security and timeliness of military and civil applications to a great extent [5,6]. With the rapid advancement of deep learning, the single-stage object detection model represented by YOLOv4 [7] has attracted considerable interests in the application of marine ship detection due to its real-time speed and relatively high accuracy. Deployment of a YOLOv4 network in unmanned airborne vehicles (UAVs) is expected to improve the development of maritime automation maintenance and early warning technology markedly, which has significant application prospects and strategic sense.

The YOLO-based networks, as typical end-to-end and concise pipelines, treat object detection as a regression problem without the region generation step [8], achieving excellent computational efficiency in many challenging benchmark datasets. However, these models find it difficult to distinguish the foreground from the background in a complex ocean scene [9], especially for detecting small-scale ships, easily resulting in false and missed alarms. Therefore, it is of urgent need to explore an effective strategy to achieve the trade-off between detection efficiency and accuracy, that is to say, it is valuable to study a more accurate model without greatly increasing the amount of calculation.

Attention mechanism [10] introduces the spatial or channel significance to the feature maps extracted by the convolutional neural network (CNN), which contributes by focusing on the characteristic properties of interested targets. Consequently, applying the attention mechanism helps to improve the representation ability of CNN through suppressing unnecessary detail and reinforcing useful features [11,12]. Inspired by this, we proposed

a novel small ship detection method termed as portable attention-guided YOLO (PAG-YOLO), which can improve the detection performance without additional computation cost. To be specific, to optimize the representation of feature information, we propose a dual attention-related module to adaptively learn the significance of features in different scales. Besides, we present a new loss function that offers better convergence performance for small object detection. Experimental results prove that the PAG-YOLO method achieves state-of-the-art detection performance under different background conditions such as heavy sea clutter and complex port compared with other advanced approaches on the public MASATI dataset.

The main contributions of the proposed method are as follows:

- A novel attention-guided module, which can significantly optimize the representation of feature information both in spatial and channel domains.
- A new loss function, which contributes to the improvement on the detection accuracy and training efficiency simultaneously.

The remainder of this paper is organized as follows. We elaborate on the technological development of small object detection and point out the defect of the YOLO-based method in Section 2. In Section 3, we give a specific description of the proposed architecture. Section 4 gives the experimental results of our method and other benchmark methods and discusses the effectiveness of the proposed method. In Section 5, we summarize and draw a conclusion.

2. Previous Related Research

Aiming at the problem of object detection, many scholars have proposed a variety of processing strategies based on the characteristics of remote sensing images [13,14]. In addition, varied methods based on deep network have been presented in recent years and achieved excellent detection performance [15,16]. These methods can be divided into two main categories: one-stage approaches and two-stage approaches. The two-stage models usually include two steps: potential region extraction and class prediction. The Region-based Convolutional Neural Network (R-CNN) [17], presented in 2014, was a significant pioneering work in the early development of two-stage detectors, which led a wave of research at that time. The R-CNN extracted image features through a selective search instead of the traditional sliding window idea and then used a classifier to predict the objects. This method improved the detection performance, but the computational cost was so large that a small dataset even needed several days to train. To solve this problem, the Fast R-CNN [18] was then proposed to optimize the training process by simplifying the redundant computation for the overlapping candidate regions. In addition, this method abandoned the idea of adopting multiple classifiers and bounding box regressors, achieving near-real-time end-to-end train speed. Based on the Fast R-CNN, the Faster R-CNN [19] model was presented in 2017, which introduced a region proposal network (RPN) to generate the possible regions, reducing the reasoning time by an order of magnitude. That same year, to enhance the performance of small object detection, another influential architecture, the feature pyramid networks (FPN) [20], was proposed to learn the features of different levels. Note that this FPN has become an essential module of the existing multi-scale detection methods. An excellent detection method needs to consider both accuracy and computational efficiency. Although the two-stage methods obtain superior accuracy, the computational speed of these methods are usually inferior to that of the single-stage approaches. For this reason, the single-stage detection methods, represented by the YOLO-based networks, are widely applied in various practical tasks, especially on the lightweight platform. In 2016, the first generation you only look once (YOLO) [21] model was developed. It transformed the detection issue into a mathematical regression problem, which greatly inspired the development of subsequent single-stage methods. That same year, the Single Shot Multi-Box Detector (SSD) [22] was presented, which provides a useful strategy to detect objects by combining features of different scales and default bounding boxes. After that, more methods were gradually added to the YOLO-

based family, and the detection rate was significantly improved. As a matter of fact, due to the research of many scholars, the YOLO-based network has developed to YOLOv4. Nevertheless, many strategies have been presented to improve the detection effect for small targets in these methods. When the detection scene elements are complex, such as islands, dense clutter, and port facilities interferences, false detection and missed detection readily occur due to the lack of a perfect preprocessing mechanism. Therefore, there is still room for improvement in the performance of small target detection. As opposed to other methods, this paper mines the global and local relationship in the image from the spatial and channel domains, optimizing the feature representation of small targets in the feature extraction stage. In addition, considering the accuracy and calculation speed of the detection model, we propose a new loss function, which enables the model to predict the shape of objects more efficiently.

3. Proposed Method

3.1. Method Overview

The proposed framework is illustrated in Figure 1. Similar to the YOLOv4, the widely used CSPDarknet53 [7] module is employed as the backbone, extracting features of different scales in the input image. Moreover, to effectively mine context information, the SPP part [23] is added over the extracting module. In the stage of integrating different scale features, the commonly used and high-quality FPN and the Path Aggregation Network (PAN) [24] are combined as the neck. Considering that the optimization of feature aggregation directly affects the final effect of the detection model, we propose a dual attention feature optimization (DAFO) module to optimize the weight distribution of each feature map. A more detailed explanation of the module is provided in Section 3.2. In the stage of classification and prediction, we propose a new function to constrain the detection head, learning the shape of the ship target faster and more accurately. The detail of the loss function is elaborated in Section 3.3.

Figure 1. Flow of the proposed method.

3.2. Dual Attention Feature Optimization

Before the prediction head in the original YOLOv4 network architecture, a convolution block is applied directly after concatenating the feature maps of different scales. It

means that the extracted features from each channel and each position contribute equally to the final detection result. But in fact, the feature maps of each channel and position can be regarded as a response of specific semantic information. Obviously, if an appropriate weight is offered to the part that represents the object characteristic, the ability of the detection model to locate the correct targets will be effectively improved. Attention mechanism can flexibly capture the global and local relationships in the input image, so as to focus on finding significant information related to the application task. Inspired by this, we introduce a dual attention mechanism to improve the original YOLO-based structure, which can adaptively optimize the weight distribution of each feature map from space and channel dimension without increasing the computation amount significantly. The optimized structure is illustrated in Figure 2. First, a channel attention module (CAM) is applied after concatenating the feature maps to establish the interdependence relation between channel mappings and to enhance the expression of specific semantic information. Then, a spatial attention module (SAM) is employed after the convolution block to gain the weight of local features and improve the ability to localize all candidate regions.

Figure 2. Dual attention module design.

The CAM generate the weight W_C in the channel dimension. Specifically, the max pooling and the average pooling operations are adopted to gather the spatial information. For the feature map $I \in \mathbb{R}^{c \times h \times w}$, c indicates the channel number, h and w indicate the height and width of the map, respectively, and the pooling vectors in the m -th channel are computed by

$$r_m^{ave} = \frac{1}{h \times w} \sum_{i=1}^h \sum_{j=1}^w I_m(i, j) \quad (1)$$

$$r_m^{max} = \max(I_m) \quad (2)$$

For the gathered spatial information, a shared fully connected layer is then applied to learn the dependencies between channels:

$$r_{FC} = f_{FCN}(r) \quad (3)$$

where $f_{FCN}(\cdot)$ indicates the fully connected network.

In order to integrate these responses, we add the feature descriptors in an element-wise way. Then, we apply the sigmoid function to obtain the weight W_C .

$$W_C = \text{sigmoid}(r_{FC}^{ave} + r_{FC}^{max}) \quad (4)$$

Before optimizing the spatial importance of different features, a convolution block is applied after the channel attention block, generating the first predictions. Then, we gain the new feature map as

$$I_1 = f_{CB}(W_C \times I) \quad (5)$$

where f_{CB} represents the convolution operator.

The SAM obtains the weight W_S in the spatial dimension. Similar to the previous step, the max pooling and the average pooling are used along the channel dimension to highlight the valid information for object detection.

$$p_{i,j}^{ave} = \frac{1}{c} \sum_{k=1}^c I_k(i, j) \quad (6)$$

$$p_{i,j}^{max} = \max(I(i, j)) \quad (7)$$

where p^{ave} and p^{max} are the pooling vectors of the map I_1 . Moreover, (i, j) indicates the coordinate of a point in the map.

For the gathered features, the dependency between the local features is extracted by using the 7×7 convolution kernel, and the weight coefficient of spatial dimension W_S is then obtained as

$$W_S = \text{sigmoid}(f_{7 \times 7}(p)) \quad (8)$$

where p is the concatenating vector of p^{ave} and p^{max} along the channel dimension, and $f_{7 \times 7}$ means the convolution process.

Finally, the feature map produced by the whole dual attention module can be calculated as

$$I_2 = W_S \times I_1 \quad (9)$$

3.3. Loss Function

To improve the model training efficiency, we proposed a new loss function, which employs the mean square error, cross-entropy, and complete intersection over union indexes (CIoU) [25] to describe the model loss synthetically. Suppose each input image is divided into $N \times N$ grids, and each grid is distributed with B bounding boxes. Similar to the previous YOLO networks, we calculate the total loss function from three aspects: bounding box position, category confidence, and class. Consequently, the loss function can be written as

$$L = L_{obj,confid} + L_{noobj,confid} + L_{position} + L_{ship} \quad (10)$$

where $L_{position}$ and L_{ship} represent the loss of the predicted box position regression and ship classification, respectively. $L_{obj,confid}$ and $L_{noobj,confid}$ represent the object confidence loss of the boxes with and without targets, respectively.

Considering the ship detection is a binary classification issue, we use the widely used and effective mean square error (MSE) index to calculate the confidence loss.

$$L_{obj,confid} = \sum_{i=0}^{N^2} \sum_{j=0}^B l_{ij}^{obj} [(Cf - Cf_{gt}) \cdot \nabla_{\sigma}(Cf)]^2 \quad (11)$$

$$L_{noobj,confid} = \sum_{i=0}^{N^2} \sum_{j=0}^B l_{ij}^{noobj} [(Cf - Cf_{gt}) \cdot \nabla_{\sigma}(Cf)]^2 \quad (12)$$

where l_{ij} is the proportion occupied by the j -th box in the i -th grid. Cf and Cf_{gt} are the confidence values obtained by the boxes. ∇_{σ} means the gradient calculation of the sigmoid function.

In order to avoid wasting time making the prediction box try to overlap with the real box, we employ the CIOU to calculate the position regression loss, which can significantly improve the convergence speed of the loss training.

$$L_{position} = 1 - IOU + \frac{\rho^2(b, b_{gt})}{l^2} + \alpha v \quad (13)$$

where b and b_{gt} define the prediction box and the real box, respectively. l is the diagonal length of the minimum bounding box which contains both two boxes. $\rho(\cdot)$ represents the Euclidean distance. αv is a penalty term that is employed to constrain the aspect ratio property and its elements are defined as

$$\alpha = \frac{v}{(1 - IOU) + v} \quad (14)$$

$$v = \frac{4}{\pi^2} (\arctan \frac{w_{gt}}{h_{gt}} - \arctan \frac{w}{h})^2 \quad (15)$$

In the aspect of category loss, in order to speed up the gradient update in the training process, we compute the cross-entropy (CE) loss to focus on the prediction probability of the correct category. The classification loss is computed as

$$L_{ship} = \sum_{i=0}^{N^2} \sum_{j=0}^B l_{ij}^{obj} [-pr(ship) \log(pr_{gt}(ship))] \quad (16)$$

where $pr(\cdot)$ and $pr_{gt}(\cdot)$ define the probability produced by the predicted classification and the ground truth.

Moreover, considering that the ship detection in remote sensing image is generally a sparse target detection problem in a large-scale scene, many grids in the image do not contain targets, and their confidence tends to be zero, thus the direct training easily leads to instability and inaccuracy of the detection model. To overcome this problem, we empirically set an additional weighting factor for each item to allocate the importance of different loss functions more reasonably. Consequently, the total loss function can be obtained as

$$L_{total} = \lambda_1 L_{obj,confid} + \lambda_2 L_{noobj,confid} + \lambda_3 L_{position} + \lambda_4 L_{ship} \quad (17)$$

where λ_1 , λ_2 , λ_3 , and λ_4 are the weights for each item. Actually, after a large number of tests and comparisons, it is found that the loss function can obtain the most satisfactory result when the parameters (λ_1 , λ_2 , λ_3 , λ_4) are set to (4, 0.5, 5, 0.5).

Through the above loss function design, it is found that the detection accuracy of the proposed model is improved; moreover, this model shows stronger robustness and faster convergence performance.

4. Experimental Results

4.1. Datasets and Evaluation Metrics

To verify the application performance of the proposed method, we carry out comparative experiments on the MASATI dataset [26]. The MASATI dataset consists of various ocean scenes, such as the pure sea, reef, and port, and collects images of single or multiple ships under different illumination conditions. The 7389 satellite images have a resolution of 512×512 pixels. We pick up the images containing a ship target which sum up to 2368 images, and randomly select 70% of them as the training set (1675 images) and the others as the testing set (711 images).

The average-precision (AP) metric is chosen to evaluate the detection performance of our method. Besides, the frames per second (FPS) index is applied to assess the real-time performance. Similar to the works in [19,27], the Intersection over Union (IOU) threshold

with ground truth bounding box is set to 0.5. AP , the area surrounded by Precision–Recall curve, is defined as follows:

$$AP = \int_0^1 precision(recall)d(recall) \quad (18)$$

where recall refers to the number of TP divided by the TP and FN (Fault Negative), and precision refers to the number of TP (True Positive) divided by the TP and FP (Fault Positive), which are calculated below:

$$precision = \frac{TP}{TP + FP} \quad (19)$$

$$recall = \frac{TP}{TP + FN} \quad (20)$$

For practical application tasks, the false alarm rate (FAR) is a key index that directly reflects whether the algorithm can respond well to the task requirements. Therefore, we also introduce this evaluation index to verify the overall effect of the proposed algorithm.

$$FAR = \frac{\text{number of detected false alarms}}{\text{number of detected candidates}} \quad (21)$$

4.2. Implementation Details

All the experiments are carried out on a workstation with 64 GB RAM and a Quadro RTX 4000 GPU using the Pytorch framework. Before the network training stage, we studied the geometric characteristics of ship targets in the dataset. Based on the prior knowledge of object position and ground-truth bounding box size, we adopted the k-means clustering method to optimize the anchor box parameters, such as height and width. After repeated iterations, we obtained the anchor parameters that meet the IOU-based loss function constraint. The anchors represented by these parameters are the customized boxes that fit the small ships in the used dataset. In the training process, we trained all the model with 300 epochs, and the batch size is set as 4. Combined with the characteristics of our dataset, the stochastic gradient descent (SGD) is used with a weight decay of 0.0005 and momentum of 0.937. Moreover, we applied the mosaic data augmentation to put four training samples into one which helps reduce the need for a large mini-batch size. The training curves of the proposed model are shown in the Figure 3.

Figure 3. The training curves of the proposed model.

4.3. Ablation Analysis

Aiming to prove the contributions of the proposed module and the new loss function, ablation experiments are carried out on the MASATI dataset. For fair comparisons, the original YOLOv4 method is selected as the baseline. Table 1 shows the experiment results of multiple image scenes on the MASATI dataset. As one can observe, when YOLOv4 is used directly, the obtained AP value is 83.5. Based on this condition, when the DAFO module is assembled on the original architecture, the AP value increases by 4.1%. When this network is further constrained by the presented loss function, the AP even increases by 7.5%. In addition, from the perspective of the FPS index, it can be found that the embedding of the DAFO module slightly reduces the calculation speed (the FPS is from 72 to 69), but the presented loss function can improve the calculation efficiency (the FPS is from 72 to 75). Overall, the proposed model does not add extra computation compared with the original YOLOv4, but significantly improves the detection accuracy.

In order to illustrate the improvement of detection effect more intuitively, Figure 4 gives the processing results of several typical scene examples. From the image results of the first row, it can be found that the original YOLOv4 method mistakenly detects an island on the sea surface as a ship, but our method effectively reduces the false alarm. For the second row, the ship in the original image has a significant wake. Compared with the YOLOv4 approach, our method reduces the interference of the wake to the target localization, offering a correct prediction box with better compactness. For the third line of the multi-target scene image, there is a missing detection in the YOLOv4 model. In contrast, our method obtains a satisfactory result.

4.4. Algorithm Performance Comparison

In order to verify the overall performance of the proposed model, comparisons are made between the proposed PAG-YOLO and several recent advanced methods. The compared methods contain the Faster R-CNN [19], DC-SPP-YOLO [27], and the recent Attention Mask R-CNN [28] model. For a fair comparison, all the compared methods adopt the same data augmentation strategy.

To fully illustrate the applicability of our model to different image scenes, we give the processing results of several typical scene conditions in Figure 5. Note that we remove the confidence label in the result images to gain a concise and clear display effect. From the original images (as shown in Figure 5a), we can see that the tested image scenes include various background elements, such as serious sea clutter, obvious ship wake, and complex port facilities. In addition, our experiments also cover the performance verification under the single target and multi-target conditions. From the experiment results of various scenes, we can find that the proposed method outperforms all the compared approaches in terms of the detection accuracy.

Specifically speaking, for a heavy sea clutter scene, as shown in the first column of Figure 5, one can see that the Attention Mask R-CNN misses the only small ship, whereas the Faster R-CNN and DC-SPP-YOLO methods show a different amount of false detection. In contrast, the proposed method can highlight the salient target and achieve effective detection. The second column shows the coast scene with the interference of shore reefs, seabed reefs, and sea waves. It can be found that all the methods achieve the success of target detection, but there is a false detection in the Attention Mask R-CNN method. Interestingly, the Attention Mask R-CNN method also uses an attention part similar to our model, but our approach fully considers the sparse distribution characteristics of the ship target and configures more reasonable weight factors, obtaining a better detection performance. For the case of a ship with strong wake, as shown in the third column, the first two methods have false detection, whereas the DC-SPP-YOLO and our method mark the correct target. Moreover, it is obvious that the box we marked fits better with the real target, reflecting the position and size of the target more accurately. In terms of the multi-target scenes, as shown in the fourth and fifth columns, it is clear that the first two methods demonstrate different degrees of miss detection, especially for the weak target and small

target. On the contrary, the DC-SPP-YOLO method shows better scale compatibility and can highlight ships with different shapes and directions.

Table 1. Results of ablation experiments.

Methods	Loss Function	AP(%)	FPS
YOLOv4	The original function	83.5	72
YOLOv4+DAFO	The original function	87.6 (+4.1)	69
YOLOv4	The proposed loss function	86.0 (+2.5)	75
YOLOv4+DAFO	The proposed loss function	91.0 (+7.5)	70

Figure 4. Comparison results of the proposed method and the YOLOv4 network: (a) input image, (b) YOLOv4, (c) our method, and (d) ground truth.

Figure 5. Detection results obtained by our method and the compared approaches: (a) original images, (b) Faster R-CNN, (c) Attention Mask R-CNN, (d) DC-SPP-YOLO, (e) the proposed method, and (f) ground truth.

In addition, Table 2 reports the quantitative results obtained by all approaches on the MASATI dataset. Clearly, the Faster R-CNN gains the lowest accuracy, and the Attention Mask R-CNN requires the longest processing time per frame. Compared with the latter

two methods, the two methods have a certain difference in term of the processing speed. The main reason is that they are two-stage detectors, and the cost of calculation is larger than that of the single-stage detectors. Besides, the performance of the DC-SPP-YOLO method shows little difference with our algorithm, but our method still has advantages in accuracy and calculation speed, achieving satisfactory detection results with a high frame rate. In addition, our algorithm obtains the lowest false alarm rate, showing more advantageous false alarm control ability than the compared methods. Consequently, through the above qualitative and quantitative results, one can conclude that the proposed method outperforms the compared advanced methods both in terms of detection performance and algorithm speed.

Table 2. Detection performance of different methods.

Methods	AP(%)	FAR(%)	FPS
Faster R-CNN	85.4	9.56	19
Attention Mask R-CNN	88.5	7.25	17
DC-SPP-YOLO	90.7	4.85	65
Proposed	91.0	4.50	70

5. Conclusions

In this paper, we propose a novel implementation of a small ship detection model named PAG-YOLO which improves the detection accuracy and does not introduce additional computational costs compared with the YOLOv4 model. As opposed to other ship detection methods, an attention-guided feature optimization module is proposed, which can adaptively reassign the weight distribution of different scale features from both space and channel dimension. Moreover, combined with the sparse distribution characteristics of ships in a remote sensing scene, a new loss function is presented to constrain the performance of the final classification, which contributes to the fast stability and accuracy of the PAG-YOLO detector. The experimental results demonstrate that the proposed architecture shows better accuracy and robustness on the public and high-quality MASATI dataset in comparison with other excellent benchmarking methods.

For ship detection and recognition based on remote sensing images, occlusion of scene elements, large-dynamic scale change and dense target distribution are all challenges for algorithm application. In the future, we plan to carry out further research on robust dense multi-scale target detection.

Author Contributions: Investigation, literature analysis, methodology, writing—original draft, validation, J.H.; funding acquisition, project administration, X.Z.; supervision, W.Z.; revising and editing, T.S., Y.C. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (61975043) and the National Key R&D Program of China(2017YFB0502902).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets for this research is released in June 2019, which is available at <https://www.iuii.ua.es/datasets/masati/>, accessed on 23 June 2021.

Acknowledgments: The authors would like to thank Antonio-Javier Gallego from the University of Alicante for providing the MASATI dataset.

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

References

- Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X.X. HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7147–7161. [\[CrossRef\]](#)
- Hu, J.; Zhi, X.; Zhang, W.; Ren, L.; Bruzzone, L. Salient Ship Detection via Background Prior and Foreground Constraint in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3370. [\[CrossRef\]](#)
- Dong, C.; Liu, J.; Xu, F. Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor. *Remote Sens.* **2018**, *10*, 400. [\[CrossRef\]](#)
- Xu, F.; Liu, J.; Dong, C.; Wang, X. Ship detection in optical remote sensing images based on wavelet transform and multi-level false alarm identification. *Remote Sens.* **2017**, *9*, 985. [\[CrossRef\]](#)
- Ji, F.; Ming, D.; Zeng, B.; Yu, J.; Qing, Y.; Du, T.; Zhang, X. Aircraft Detection in High Spatial Resolution Remote Sensing Images Combining Multi-Angle Features Driven and Majority Voting CNN. *Remote Sens.* **2021**, *13*, 2207. [\[CrossRef\]](#)
- Tan, Z.; Zhang, Z.; Xing, T.; Huang, X.; Gong, J.; Ma, J. Exploit Direction Information for Remote Ship Detection. *Remote Sens.* **2021**, *13*, 2155. [\[CrossRef\]](#)
- Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
- Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13039–13048.
- Hu, J.-M.; Qiao, K.; Zhi, X.-Y.; Zhang, Y.; Gong, J.-N.; Chen, W.-B. Influence of complex environment on the detectability of weak and small aerial target under space-based observation mode. *J. Infrared Millim. Waves* **2019**, *38*, 351–357.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
- Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
- Ju, M.; Luo, J.; Wang, Z.; Luo, H. Adaptive feature fusion with attention mechanism for multi-scale target detection. *Neural Comput. Appl.* **2021**, *33*, 2769–2781. [\[CrossRef\]](#)
- Jiang, S.; Zhi, X.; Zhang, W.; Wang, D.; Hu, J.; Chen, W. Remote sensing image fine-processing method based on the adaptive hyper-Laplacian prior. *Opt. Lasers Eng.* **2021**, *136*, 106311. [\[CrossRef\]](#)
- Jiang, S.; Zhi, X.; Zhang, W.; Wang, D.; Hu, J.; Tian, C. Global Information Transmission Model-Based Multiobjective Image Inversion Restoration Method for Space Diffractive Membrane Imaging Systems. *IEEE Trans. Geosci. Remote Sens.* **2021**. [\[CrossRef\]](#)
- Yang, Y.; Pan, Z.; Hu, Y.; Ding, C. CPS-Det: An Anchor-Free Based Rotation Detector for Ship Detection. *Remote Sens.* **2021**, *13*, 2208. [\[CrossRef\]](#)
- Yang, Z.; Tang, J.; Zhou, H.; Xu, X.; Tian, Y.; Wen, B. Joint Ship Detection Based on Time-Frequency Domain and CFAR Methods with HF Radar. *Remote Sens.* **2021**, *13*, 1548. [\[CrossRef\]](#)
- Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences (2014). *arXiv* **2017**, arXiv:1404.2188.
- Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
- 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.
- Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–16 July 2017; pp. 2117–2125.
- 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.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–21 June 2018; pp. 8759–8768.
- Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
- Gallego, A.J.; Pertusa, A.; Gil, P. Automatic ship classification from optical aerial images with convolutional neural networks. *Remote Sens.* **2018**, *10*, 511. [\[CrossRef\]](#)
- Huang, Z.; Wang, J.; Fu, X.; Yu, T.; Guo, Y.; Wang, R. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. *Inf. Sci.* **2020**, *522*, 241–258. [\[CrossRef\]](#)
- Nie, X.; Duan, M.; Ding, H.; Hu, B.; Wong, E.K. Attention mask R-CNN for ship detection and segmentation from remote sensing images. *IEEE Access* **2020**, *8*, 9325–9334. [\[CrossRef\]](#)