Multi-scale Target Detection Algorithm of Optical Remote Sensing Image Based on Improved YOLOv8

Haoyu Wang[1], Haitao Yang [2,*], Jinyu Wang [1], Xixuan Zhou[1], Yifan Xu [1], Honggang Zhang [1]

[1] Graduate School, Space Engineering University, Beijing, China, 2018205188@qdu.edu.cn

[2] Space Engineering University, Beijing, China, yanghtt@126.com

* Correspondence: yanghtt@126.com

Keywords: Target Detection, Remote Sensing Image, YOLOv8, SPPF, Loss Function.

Abstract

With the progress of remote sensing sensors, the quality of optical remote sensing image is significantly improved, and target detection on it can extract rich feature information. However, due to the characteristics of remote sensing image with various target sizes and a large proportion of the number of small targets, increasing the difficulty in target detection for it. In response to this challenge, this paper proposes an improved YOLOv8 algorithm for multi-scale target detection of optical remote sensing images. First, we propose a PSPPF module, which improves the model's ability to adapt to different data distributions; Second, DSConv is introduced into the Backbone structure of YOLOv8 to reduce the complexity of the network while maintaining the performance of model detection; Finally, we replace the original loss function CIoU with MPDIoU to improve the localization accuracy of the prediction box. Applying the improved algorithm to the public dataset NWPU VHR-10, the mAP value of the our algorithm is 95.1%, which is 3.0% higher than that of the original YOLOv8, indicating that the proposed algorithm is able to effectively detect multi-scale targets in optical remote sensing images.

1. Introduction

Optical remote sensing images are acquired by optical sensors, which contains information about objects and changes on the Earth's surface. Through target detection of these images, researchers can analyze the characteristics and movement of different targets without touching or approaching them. Therefore, remote sensing image target detection has been widely used in urban planning, geological survey, hydrological analysis, national defense and many other fields (Wang et al. 2024). With the development of aerospace technology and the advancement of remote sensing imaging equipment, the quality of optical remote sensing images has been significantly improved. However, the complexity of the information contained in the image has also increased markedly, and there are generally characteristics such as diverse target sizes, dense target distribution and high proportion of background information, which increases the difficulty of doing target detection on it.

Traditional target detection is mostly achieved by feature-based and segmentation-based methods, such as DPM features and Harr features, which have provided a variety of generalized algorithms for target detection tasks for a long time due to their low computational effort and fast processing speeds. In recent years, with the increase in the complexity of optical remote sensing images, the traditional methods are gradually unable to meet the needs of the current multi-scale target detection tasks in terms of detection accuracy and speed. Meanwhile, because of the maturity of deep learning networks and their wide application in the field of image processing, deep learning algorithms with higher detection accuracy and stronger generalization have further promoted the development of target detection methods, and have gradually become the mainstream way. Current target detection algorithms based on deep learning can be mainly categorized into two types: two-stage detection and one-stage detection. Two-stage detection methods, such as R-CNN (Girshick et al. 2014), Fast R-CNN (Girshick et al. 2015), and Faster R-CNN (Ren et al. 2015), need to select candidate regions on the image firstly, and analyze these regions in order to arrive at the classification and localization of targets, which have a higher accuracy but slower speed. In order to improve the detection speed, one-stage detection algorithm represented by the YOLO (Redmon et al. 2016) series and SSD (Liu et al. 2016) series has appeared. This type of method omits the candidate region generation stage, and after the image features are extracted with neural networks, the category and position coordinates of the target are directly obtained through regression analysis, which is suitable for task scenarios such as high-volume detection and real-time monitoring. Although the YOLO series have better performance in many application scenarios, the accuracy of using them detecting targets directly is relatively low because of the complexity of optical remote sensing images. In order to improve the detection performance of targets in images, many researchers have improved the YOLO series of algorithms. (Xie et al. 2023) improves the YOLOv4 algorithm by adding the adaptive Spatial Feature Fusion structure into the feature enhancement network, and at the same time optimizes the Spatial Pyramid Pool structure, and adopts the adaptive weight parameter to fuse the multi-scale feature information to improve the detection accuracy of the remote sensing targets; (Li et al. 2023) introduces a novel YOLOv5-based network named RSI-YOLO, they utilized channel attention and spatial attention mechanisms to enhance neural network fusion of features, and improved the original network multi-scale feature fusion structure based on the PANet structure into a weighted bi-directional feature pyramid structure to achieve more efficient and richer feature fusion; (Liu et al. 2023) presents an innovative approach called YOLOv8-SnakeVision. The method introduces Dynamic Snake Convolution, Context Aggregation Attention Mechanism and Wise-loU strategy in the YOLOv8 framework to improve the target detection performance. It not only enhances small object

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-649-2024

detection but also strengthens the ability to recognize multiple targets; (Yi et al. 2023) proposed LAR-YOLOv8. This algorithm enhances the local module in the feature extraction network by utilizing the dual-branch architectural attention mechanism, and proposes the RIOU loss function, which avoids the failure of the loss function and improves the consistency of the shape of the prediction box with the ground truth box.

Based on the inspiration of the above work, we propose a multi-scale target detection algorithm for optical remote sensing images based on improved YOLOv8 and has achieved good results on the NWPU VHR-10 (Gong et al. 2014). The main contributions of this paper can be summarized as follows:：

(1)The PSPPF module is proposed to improve the sensitivity of the network to image detail information and the detection of multi-scale targets.

(2)Add DSConv to the Backbone structure of YOLOv8, to reduce the number of computational parameters of the network and accelerate the training speed while ensuring the model has a high detection accuracy.

(3)Use MPDIoU instead of the original loss function CloU, in order to enhance the localization accuracy of the prediction box and improve the accuracy of target detection.

2. YOLOv8 Network

YOLOv8 is a one-stage target detection algorithm open-sourced by Ultralytics, and its specific network structure is shown in Figure 1.

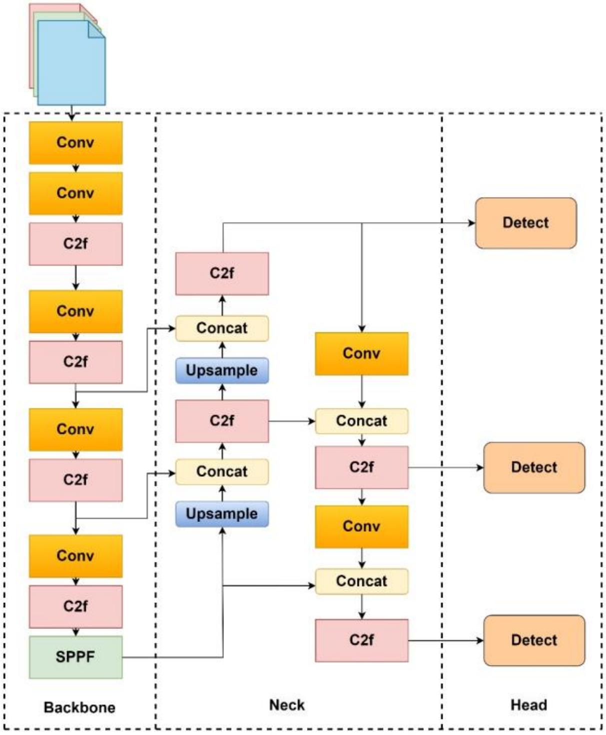


Figure 1. The structure of YOLOv8.

The YOLOv8 algorithm model is mainly composed of three parts: Backbone, Neck and Head. Before the model training starts, the input side will first preprocess the images fed into the network through Mosaic data enhancement, adaptive image scaling and grayscale padding. The Conv, C2f and SPPF modules in the Backbone extract image features by convolution and pooling and input them to the Neck. Neck is designed based on the PAN (Path Aggregation Network), which fuses feature maps with different scaling scales through up-sampling, down-sampling and splicing. Head is composed of anchor-free decoupled head structure, which realizes positive and negative sample matching and loss calculation.

There are five different pre-trained models for YOLOv8, which are YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. YOLOv8n is the smallest of all the models, and the others are obtained by increasing the width as well as the depth of the training based on it.

3. Improvement of YOLOv8

3.1 PSPPF Module

The SPPF module is located in the last layer of the Backbone and it utilizes the output of the previous layer as the basis for processing. SPPF can effectively fuse the global feature information from the feature extraction network to improve the efficiency and accuracy of the model. However, the traditional SPPF module uses the ReLU (Glorot et al. 2011) as the activation function. ReLU has a zero output value in the negative input region, which leads to the fact that when a neuron learns negative weights during the training process, the neuron will never be activated, and will be unable to update its weights, resulting in a partial failure of the network. To solve this problem, we propose a PSPPF module that can better adapt to different data distributions by introducing the PReLU (He et al. 2015) function into the SPPF. The specific structure of PSPPF is shown in Figure 2.

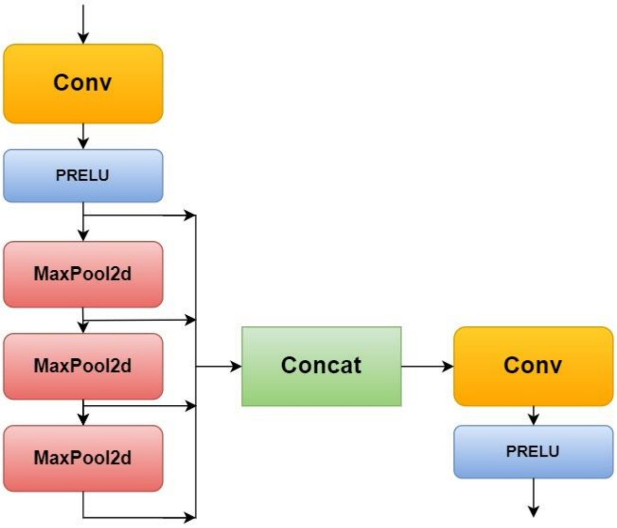


Figure 2. The structure of PSPPF

where PReLU is specified as follows：

$$f(x_i)=\begin{cases} x_i & x_i > 0 \\ a_i x_i & x_i \le 0 \end{cases} \quad (1)$$

Where a_i is a learnable parameter. PReLU has different slopes in the negative input region, which enables the PSPPF module to learn different feature representations based on the input data and improves its robust performance. Meanwhile, relative to the asymmetric property of ReLU, PReLU can better balance the processing of positive and negative inputs and improve the model representation.

3.2 Add DSConv into the Backbone

DSConv (Nascimento et al. 2019) is a sensitive quantized convolutional operator, which can effectively improve the speed and reduce the memory consumption of convolutional neural networks while guaranteeing the training effect, and its specific computational process is shown in Figure 3.

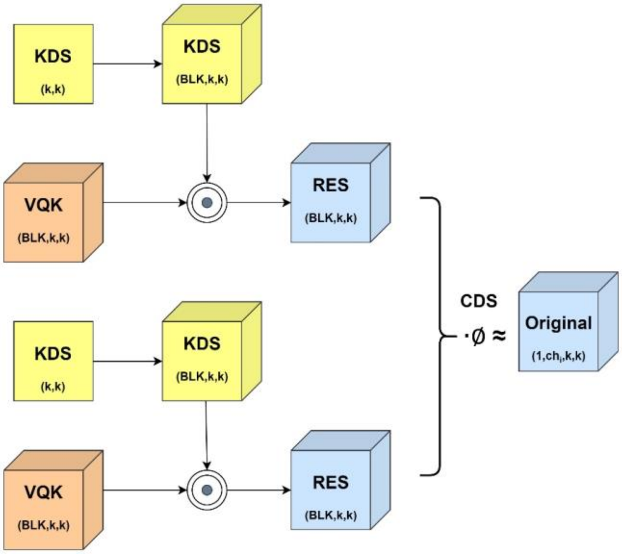


Figure 3. The structure of DSConv

When doing the convolution operation, DSConv splits each tensor into two parts: The first part is called the Variable Quantized Kernel (VQK), which has the same tensor size as that obtained by regular convolutional computation and consists of integer values. The second part consists of the Kernel Distribution Shift(KDS) and the Channel Distribution Shift (CDS). During pre-training of the network, DSConv first divides the weight tensor by depth into blocks of length B, each sharing a floating-point value, then quantizes all the blocks using the block floating point (BFP), and finally multiplies the quantized blocks with the integer values of the first part of the tensor, and multiplies the obtained results by their respective scales again in order to achieve the distribution of the individual blocks in the correct range. The memory savings per tensor for calculations by using DSConv are：

$$p = \frac{b}{32} + \frac{\left\lceil \frac{C_i}{B} \right\rceil}{C_i} \tag{2}$$

where is the number of input channels and is a customizable hyperparameter. In this paper, DSConv is added to the Backbone structure to reduce the number of parameters during training.

3.3 Improvement of the Loss Function

In the target detection task, the loss function plays an important role in the localization accuracy of the detection results. Therefore, choosing the appropriate loss function according to different task requirements can improve the learning effect and robustness of the model. The traditional IoU loss function evaluates the detection results by measuring the similarity between the predicted frame and the actual frame. In recent years, many different loss functions have been derived based on IoU, which improve the defects of the original IoU loss function from different aspects, and the most representative methods are GIOU, DIOU and CIOU.

YOLOv8 adopts CIoU as the loss function, which integrates three important geometric factors including overlap region, centroid distance and aspect ratio, resulting in improved localization of the prediction frame. However, when the prediction box has the same aspect ratio as the groundtruth box but has completely different width and height values, it may prevent the model from optimizing the similarity efficiently, and therefore there is still a lot of room for improvement in CIoU. To solve this problem, we choose MPDIoU (Siliang et al. 2023) to replace the CIoU in the original network, which is represented in Figure 4.

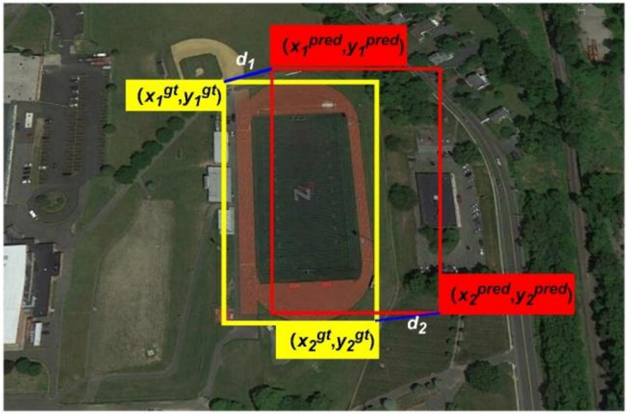


Figure 4. Schematic calculation of MPDIoU

Compared with the existing loss functions, MPDIoU takes the minimum point distance into account in addition to the relevant factors such as centroid distances, width and height deviations to recharacterize the loss function and reduce the total degrees of freedom of the loss function. The formula of L_{MPDIoU} is shown as follows:

$$L_{MPDIoU} = 1 - MPDIoU \tag{3}$$

$$MPDIoU = IoU - \frac{d_1^2 + d_2^2}{h^2 + w^2} \tag{4}$$

$$d_1^2 = (x_1^{pred} - x_1^{gt})^2 + (y_1^{pred} - y_1^{gt})^2 \tag{5}$$

$$d_2^2 = (x_2^{pred} - x_2^{gt})^2 - (y_2^{pred} - y_2^{gt})^2 \tag{6}$$

In cases such as non-overlapping centroids, MPDIoU can promote the prediction box to be closer to the groundtruth box, which solves the limitation of the CIoU. Therefore, in this paper, we use MPDIoU as the loss function of our network, which is able to stabilize the convergence of the model and improve the accuracy of the detection of multi-scale targets.

4. Experimentation and Analysis

4.1 Dataset and experimental environment

In order to test the effectiveness of the improved YOLOv8 algorithm, we choose NWPU VHR-10 (Gong et al. 2014) for training and evaluating the model. NWPU VHR-10 was released by Northwestern Polytechnical University in 2014, and contains a total of 800 high-resolution aerial images of 3651 targets in 10 categories, including airplane, ship, storage tank, baseball diamond , tennis court, basketball court, ground track field, harbor, bridge, and vehicle. The specific distribution of objectives by category is shown in Figure 5.

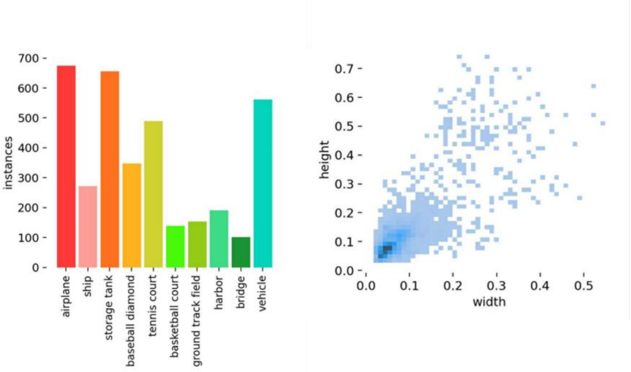


Figure 5. The number and size distribution of labels for each category

This experiment is based on the PyTorch deep learning framework and uses a single NVIDIA GeForce GPU 3090 for model training. Before starting the training, epochs were set to 300, batch_size to 16, initial learning rate to 0.01, image_size to 640×640, and the input images are preprocessed using Mosaic image enhancement. The specific experimental setting is shown in Table 1

Item	Name
OS version	Windows11
CPU	AMD Ryzen 9 5900X
GPU	NVIDIA GeForce RTX 3090
RAM	32 GB
DL framework	PyTorch (1.13.1)
Interpreter	Python (3.10)
CUDA version	CUDA (11.7)

Table 1. settings of the experimental environment.

4.2 Evaluating Indicator

In order to facilitate the evaluation of the performance of the improved model, we choose the P (Precision), the R (Recall) and the mAP (mean Average Precision) as the evaluation metrics in this experiment. The specific formulas for P and R are as follow:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

where TP (True Positive) denotes the number of samples that were judged to be positive and were in fact positive; TN (True Negative) denotes the number of samples that were judged to be negative and were in fact negative; FP (False Positive) denotes the number of samples that were judged to be positive but were in fact negative; and FN (False Negative) indicates the number of samples that were judged to be negative but were in fact positive. According to the values of P and R, the P-R curve can be plotted, and the area enclosed by the P-R curve indicates the Average Precision (AP) of a single category in the test sample, which is calculated as follows：

$$AP = \int_0^1 PR\mathrm{d}R \tag{9}$$

In target detection containing multiple categories of samples, mAP is obtained by summing the detection accuracies of each target to obtain the mean value.

$$mAP = \frac{\sum_{k=1}^{n} AP_K}{n} \tag{10}$$

4.3 Ablation study

In order to verify the effectiveness of our proposed three improved modules, we designed ablation experiments for the improved algorithms, and the specific experimental results are shown in Table 2, where Base denotes the original YOLOv8s model, "√" denotes the added module, and "× " indicates that the module is not added.

Base	PSPPF	DSConv	MPDIoU	mAP@0.5%
√	×	×	×	92.1
√	√	×	×	92.3
√	×	√	×	94.1
√	×	×	√	94.3
√	√	√	×	94.4
√	√	×	√	94.8
√	√	√	√	95.1

Table 2. the results of ablation study.

As seen from the results in the Table 2, the mAP improves by 0.2% when the PSPPF module is added to YOLOv8s alone, and by 2.0% when DSConv is added to Backbone alone. When the MPDIoU loss function is added, mAP improved by 2.2%. When all three modules are added, map improves by 3.0%, demonstrating that all of our proposed improvements contribute to improving the mAP value to varying degrees and lead to a significant improvement in the overall detection accuracy of the model.

The curves of the experimental results comparing the improved algorithm with the original algorithm are shown in Figure 6. From Figure 6, it can be seen that the mAP values of the improved algorithm are significantly higher than those of the original YOLOv8 algorithm as the number of training epochs increases.

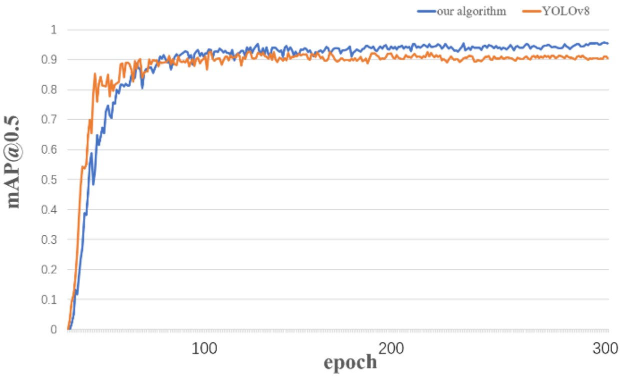
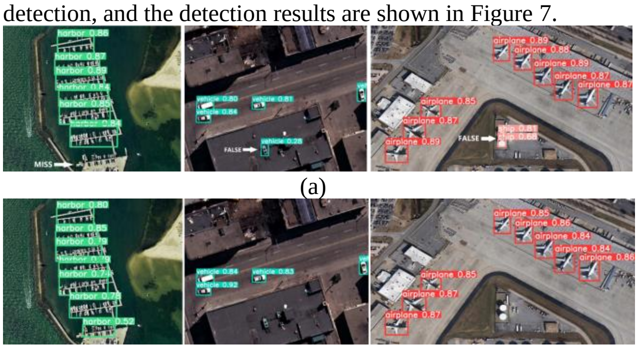


Figure 6. Comparison of mAP@0.5 curves before and after improvement.

In this paper, we select three targets of different scales which are harbor, vehicle and airplane, and compare the detection effect with the original YOLOv8 network to verify the effectiveness of the proposed algorithm for multi-scale target detection, and the detection results are shown in Figure 7.



(a)

(b)

Figure 7. Comparison of detection results between original YOLOv8(a) and improved YOLOv8(b).

For harbor, the original YOLOv8 model misses one, and the improved model detects all of them; for vehicle, the original network model misdetects one, and the improved model detects all of them; for images containing airplanes, the original network model misdetects two, and the improved model detects all of them; and for all the three types of different scales of targets, the improved algorithm's detection accuracies are higher than that of the original algorithm, which proves that our model has significantly improved detection effectiveness of multi-scale targets compared with the original YOLOv8.

4.4 Comparison test

In order to further verify the effectiveness of the improved algorithm, we choose some popular algorithms in the field of optical remote sensing image target detection in recent years to compare with our proposed algorithm, and the specific comparison algorithms and experimental results are shown in Table 3.

Algorithm	img_size	epoch	mAP@0.5%
SSD	640×640	300	81.2
Faster R-CNN	640×640	300	84.7
YOLOv4	640×640	300	89.5
YOLOv5	640×640	300	90.8
YOLOv7	640×640	300	92.7
YOLOv8	640×640	300	92.1
Our algorithm	640×640	300	**95.1**

Table 3. the results of comparison test.

From the data in Table 3, it can be seen that under the same input image size and number of training epochs, our algorithm obtains a higher mAP index compared to other algorithms, which proves that the improved algorithm has a more accurate detection effect for multi-scale targets.

5. Conclusion

For the problems of high detection difficulty and low accuracy caused by multi target scales and high background complexity in optical remote sensing images, this paper proposes a multi-scale target detection algorithm for optical remote sensing images based on the YOLOv8 by proposing a PSPPF module, adding DSConv to the Backbone structure and using MPDIoU to replace the CIoU loss function in the original network. Both ablation and control experiments demonstrate that the detection effect of our proposed algorithm is better than the original YOLOv8 algorithm and other typical target detection algorithms, and it has more excellent detection performance for multi-scale targets in optical remote sensing images. In the subsequent tasks, we plan to continue to enhance the generalization ability of the model, so that it can also achieve excellent detection results for small or ultra-small targets in visible remote sensing images.

References

Wang H, Yang H, Chen H, et al. A Remote Sensing Image Target Detection Algorithm Based on Improved YOLOv8[J]. Applied Sciences, 2024, 14(4): 1557.

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.

Girshick, R. Fast R-CNN. arXiv 2015. https://doi.org/10.48550/arXiv.1504.08083

Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Pro-ceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; pp. 21–37.

Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington DC, USA, 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 Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.

Xie T, Han W, Xu S. YOLO-RS: A More Accurate and Faster Object Detection Method for Remote Sensing Images[J]. Remote Sensing, 2023, 15(15): 3863.

Li Z, Yuan J, Li G, et al. RSI-YOLO: object detection method for remote sensing images based on improved YOLO[J]. Sensors, 2023, 23(14): 6414.

Liu Q, Liu Y, Lin D. Revolutionizing Target Detection in Intelligent Traffic Systems: YOLOv8-SnakeVision[J]. Electronics, 2023, 12(24): 4970.

Yi H, Liu B, Zhao B, et al. Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023.

Gong Cheng, Junwei Han, Peicheng Zhou, Lei Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS Journal of Photogrammetry and Remote Sensing, 98: 119-132, 2014.

Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 315-323.

He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

Nascimento M G, Fawcett R, Prisacariu V A. Dsconv: Efficient convolution operator[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5148-5157.

Siliang M, Yong X. Mpdiou: a loss for efficient and accurate bounding box regression[J]. arXiv preprint arXiv:2307.07662, 2023.