SNG-YOLOX: Non-obvious remote sensing target detection based on enhanced YOLOX

Yuan Mei (meiyuan2551161628@163.com)
School of Electronics and Information Engineering, Lanzhou Jiaotong University
Kaijun Wu
School of Electronics and Information Engineering, Lanzhou Jiaotong University
Zehao Xu
School of Electronics and Information Engineering, Lanzhou Jiaotong University
Hongquan Shan
School of Electronics and Information Engineering, Lanzhou Jiaotong University
Mengsi Wang
School of Electronics and Information Engineering, Lanzhou Jiaotong University

Article

Keywords: remote sensing image, upgrade YOLOX, target detection, non-explicit target

Posted Date: April 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1563546/v1

SNG-YOLOX : Non-obvious remote sensing target detection based on enhanced YOLOX

Yuan Mei[1], Kaijun Wu[1,*], Zehao Xu[2], Hongquan Shan[3], and Mengsi Wang[4]

[1,2,3,4]School of Electronics and Information Engineering, Lanzhou Jiaotong University, China
[*]meiyuan2551161628@163.com

Abstract

In the problem of remote sensing target detection, due to the large differences in size and shape of the target, uneven distribution of the target, and strong background interference in remote sensing images, it is difficult to identify and loca te the detection algorithm. To solve this problem, this paper starts from YOLOX. Firstly, self-calibration convolution is used to replace the deep convolution module in CSPLayer in YOLOX to enhance the context-aware ability of the model and expand the receptive field of the model. Then, in order to solve the interference of local information caused by too much consideration of the surrounding context information in the self-calibration convolution, the normalized attention module is introduced at the end of the self-calibration convolution to enhance the attention to local and non-obvious information. Finally, considering the defects of IOU in boundary box regression, GIOU loss is introduced to replace the boundary box regression. Experiments on RSOD remote sensing dat a set show that the improved algorithm improves the detection performance of 7.87 % mAP compared with YOLOX on this data set, and improves 22.03 % AP on non-obvious overpass images. In addition, compared with the current advanced detection algorithm, the improved algorithm has advanced detection performance.

Keywords: remote sensing image; upgrade YOLOX; target detection; non-explicit target

1 Introduction

Target detection of remote sensing images is one of the important tasks of remote sensing interpretation. With the development of deep convolution neural network, its research and application have been greatly promoted[1]. However, due to the large difference in target size and shape, uneven target distribution, and strong background interference in remote sensing images, the difficulty of

target detection algorithm in remote sensing target location is increased. Visual attention is a perception mechanism of human brain, which can better achieve attention to specific goals. Visual attention is mainly divided into channel attention and spatial attention, which can be reflected in the attention to the context information and local information of specific targets in a specific scene[2].

The introduction of visual attention mechanism has brought the gospel to solve the problem of remote sensing image target detection. In addition, the network structure, training strategy and data processing method also have great influence on the detection performance[3]. There have been many novel algorithms in target detection, which can be divided into five aspects : the method based on multi-scale prediction, the method based on data enhancement, the method based on feature resolution enhancement, the method based on context information and the method based on new backbone network and training strategy.

The proposed feature pyramid networks (FPN) structure greatly improves the performance of target detection[18]. After embedding FPN into the regional candidate network of Faster R-CNN, the average accuracy of small target detection of Faster R-CNN on COCO dataset is improved by 17.5 %. PANet further improves the performance of target detection, and increases the AP index in the COCO2017 target detection competition by about 3 %[4].

The consideration of context information can also improve the model performance to a certain extent. With the deepening of the layers of deep neural network, the actual receptive field of the model is gradually reduced, which will eventually lead to a certain misjudgment in the detection of complex targets. CoupleNet obtains the surrounding context information of the target in the image by doubling the feature map corresponding to ROI, and reduces the probability of misjudgment[5]. Finally, the Aps index of CoupleNet on the COCO2015 test set was 2.6 % higher than that of R-FCN[26].

The design of backbone network structure also has important influence on target detection. Due to the increase of network depth, the down-sampling process is bound to cause a large loss of original information. For example, the design of DetNet residual structure better realizes the preservation of large feature map scale, and has clearer edge information, which is conducive to the identification of small targets[40]. The detection accuracy is improved from 35.8 % of ResNet-50 to 40.2 % on MS COCO dataset[6].

YOLOX algorithm[7] is a new algorithm based on YOLOv3 structure, which adopts data enhancement, residual backbone network, feature pyramid, decoupling detection head, SimOTA label allocation, anchorless detection box and other strategies. The detection performance of YOLOXL on COCO data set is higher than that of YOLOv5-L[8]. In this paper, it is found through experiments that the detection performance of YOLOX-S on RSOD remote sensing data set is much higher than that of other target detection algorithms.

However, the experiment found that the current detection performance of YOLOX-S on non-obvious remote sensing images was far superior to other target detection algorithms, but it was still not high enough. Therefore, this paper analyzes the context information and visual attention of features, and

constructs the SNG-YOLOX model. The experimental results show that the SNG-YOLOX algorithm has 7.87 % mAP improvement compared with YOLOX in remote sensing target detection on RSOD remote sensing data set. Among them, SNG-YOLOX has an increase of 22.03 % AP, especially in non-obvious remote sensing images (overpass images).

Our contributions are as follows:

- The self-calibration convolution is used to replace the deep convolution module in the original model CSPLayer, which enhances the perception range of the network, takes into account the target context information, and greatly improves the attention of the model on non-dominant features.
- The normalized convolution module is introduced to process the CSPLayer input information, which improves the detection accuracy on various targets and further enhances the extraction of non-dominant features.
- The generalized intersection joint loss function is used for boundary box regression to improve the detection performance of the model again.
- The SNG-YOLOX receptive field analysis was carried out by using Grad-CAM++. Combined with the target detection results, the effectiveness of the model construction strategy in this paper was verified.

2 Related work

2.1 Target detection

YOLO series[8, 9, 10, 11, 12] single-stage target detection algorithms have been committed to pursuing the balance between speed and accuracy, which has important theoretical research and industrial application value. In the design process of detection algorithm, CSPDarknet[13], EfficientNet[14], Swin Transformer[15] backbone network plays a key role in feature extraction, has stronger detail retention ability, better attention to context information of the backbone network has attracted attention. The application of CSPNet[16] structure in backbone network also enhances the feature extraction ability of backbone network. The design of residual structure reduces the problems of gradient disappearance and model training caused by too deep network, but too deep network structure is bound to reduce the deep network in the feature perception area. Therefore, the introduction of attention mechanism makes the model focus on effective information, and inhibition of invalid information can solve this problem to a certain extent. TPH-YOLOv5[17] integrated CBAM[2] attention mechanism and Transformer[19] detection head , and used four detection heads to construct the network, which improved the detection performance of the network on small targets. However, in TPH-YOLOv5, CBAM is mainly used for feature pyramid fusion, and does not enhance the feature extraction ability of the main network. Recently, open-minded technology has launched YOLOX[7].

While retaining the advantages of YOLO series algorithms, its detection accuracy has been further improved. The use of some novel model construction strategies is instructive. However, YOLOX still uses the CSPDarknet53 backbone network structure, and has not made great changes, that is, the overall feature extraction ability of the network has not been substantially improved.

Considering that YOLOX has good detection performance, this paper selects YOLOX as the basis of remote sensing target detection research, and improves the main network structure of YOLOX, aiming at improving the feature extraction ability of the model and enhancing the detection ability of the model for non-obvious remote sensing targets.

2.2 Context information attention

In recent years, some novel structures have achieved remarkable results[20, 21]. ResNets constructs a deep network model by introducing residual connections and using batch normalization operations. ResNeXt[22] and Wide ResNet[23] extend ResNet by grouping or increasing the width of convolution to extract and retain rich feature information. DPN uses residual connection and dense connection to construct complex feature representation. SENet[24] introduces squeeze and excitation operations to enhance the interdependence between channels. GENet[25], CBAM[2], GCNet[27] and others introduce channel attention mechanisms or design advanced attention modules to further enhance global and local considerations of the original information. Such as the above network to build long-distance dependence can effectively enhance the understanding and attention of context information model, better achieve the extraction and retention of feature information. Different from the above methods, Self-Calibrated Convolutions (SCNet)[28] is considered from the perspective of convolution filter design. Multi-branch structure is used to extract and combine the feature information, which better retains the context information of the original feature and has a larger receptive field.

Therefore, this paper uses SCNet structure for multi-angle attempts, and found that when SCNet is used in the residual network structure, the effect is better.

2.3 Attention mechanism

Attention mechanism plays an important role in text processing and image processing, which enables the model to pay more attention to important information and ignore secondary information, so as to improve the performance of the model. SENet[24] integrates spatial information into channels to suppress unimportant features to improve network performance. BAM[29] constructed separate space and channel sub-modules in parallel. CBAM[2] realized the attention to the detection target by serial spatial and channel attention, which is equivalent to considering the context information and local information of the target. TAM[30] considers the correlation between dimensions by rotating feature maps. However, these methods ignore the weight information from training

adjustment. For this reason, Yichao Liu et al. proposed Normalization based Attention Module (NAM)[31] to reconsider the spatial and channel information from the perspective of weight, and realized the comprehensive consideration of salient features and non-obvious features.

To this end, this paper uses the NAM module to consider the surrounding context information too much after the introduction of SCNet, resulting in the interference of local information. It is found through experiments that the performance of NAM on the output side of SCNet is better, and the detection accuracy of each category of targets is improved.

3 SNG-YOLOX

3.1 Overview of YOLOX

YOLOX is an empirical improvement of the YOLO series algorithms by adopting advanced strategies such as decoupling detection head, SimOTA label allocation strategy and anchorless detection box. The detection effect of YOLOX-L exceeded YOLOv5-L on the COCO data set, winning the first place in the streaming media perception challenge (CVPR2021 self-driving seminar).

Self-Calibrated Convolutions (SCNet), Normalization-based Attention Module (NAM) and Generalized Intersection over Union Loss (GIOU Loss) are used to improve the YOLOX algorithm and named SNG YOLOX. SNG-YOLOX-S was used for experiments on the RSOD remote sensing data set, and the mAP was improved by 7.87

3.2 SNG-YOLOX

On the whole, this paper follows the original network structure of YOLOX, but improves the residual module (CSPLayer). The network architecture of SNG-YOLOX is shown in Figure 1.

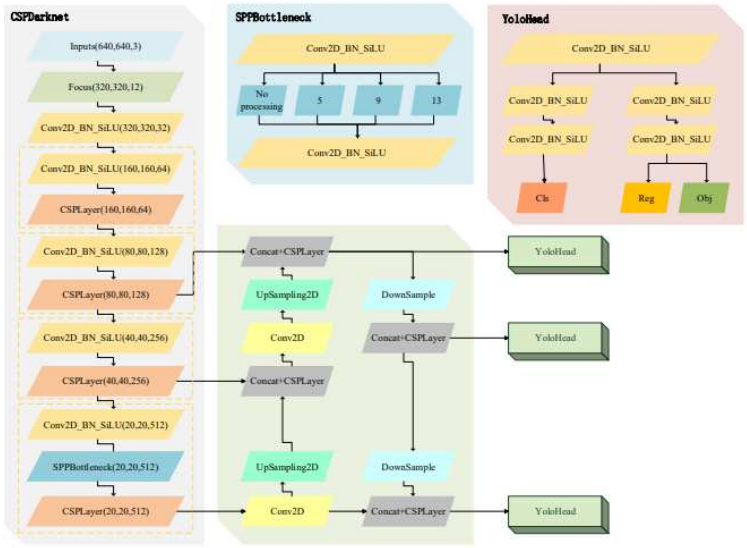
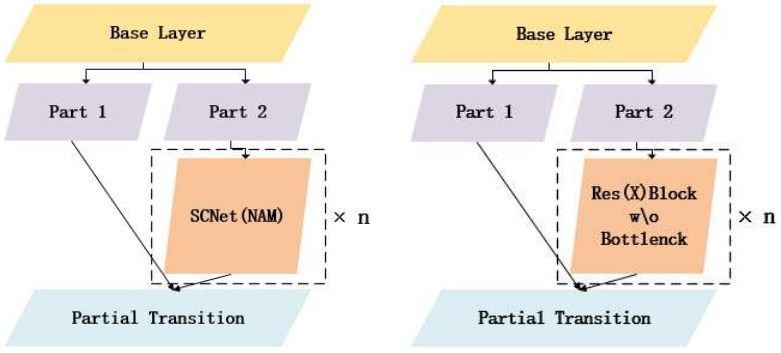


Figure 1: SNG-YOLOX network architecture diagram

In YOLOX, CSPLayer module is connected by two branches, the specific structure is shown in Figure 2. The residual structure is used to connect the model, which can ensure the model in shallow and deep feature extraction, and alleviate the problem of gradient disappearance caused by increasing the depth of the model.



(a) SNG-YOLOX (b) YOLOX

Figure 2: CSP module structure diagram

In this paper, self-calibration convolution (SCNet) and normalized attention module (NAM) are used as deep convolution parts in residual network, which can better alleviate the problem of gradient disappearance and extract

effective features. Compared with the original convolution module, the self-calibration convolution can further increase the convolution receptive field, and better consider the context information in feature extraction. The normalized attention module focuses on the attention to the effective information in the original image, which is conducive to the positioning of the target.

3.3 Self-calibration convolution

Different from the traditional 2D convolution, self-calibration convolution[28] uses multi-branch structure for different feature extraction and attention, and the specific structure is shown in Figure 3. Using multi-branch convolution, the self-calibration convolution module can not only adaptively focus on the surrounding context information, but also pay attention to the channel dependence, so it can effectively expand the convolution receptive field. In addition, the multi-branch structure can realize the multi-scale encoding of the original information, which plays an auxiliary role in the positioning of the target.

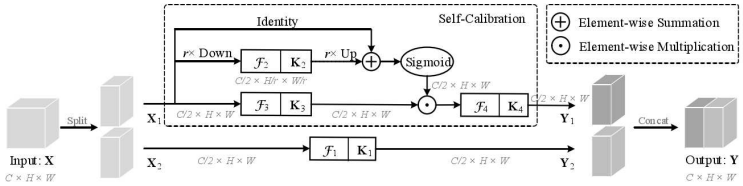


Figure 3: Self-calibration convolution structure diagram

3.4 Normalized attention mechanism

Attention mechanism helps the model better focus on important information and ignore the interference of secondary information. In remote sensing target detection, it is often difficult to distinguish between target and background. Based on the normalized attention module[31], the channel and spatial attention module are redesigned by using the module integration of CBAM. Using the contribution factor of weight to improve the attention mechanism can better focus on non-obvious indigenous information. The channel attention module and spatial attention module are shown in Fig. 4 and Fig. 5.

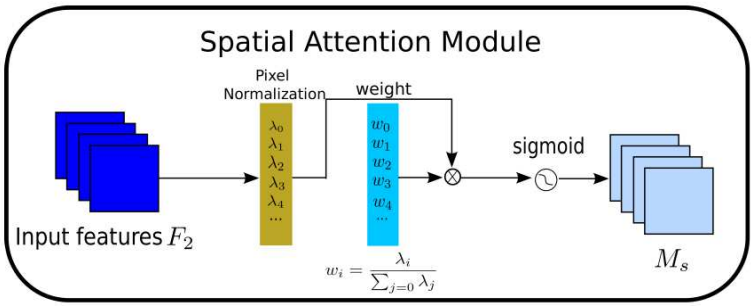


Figure 4: NAM channel attention module

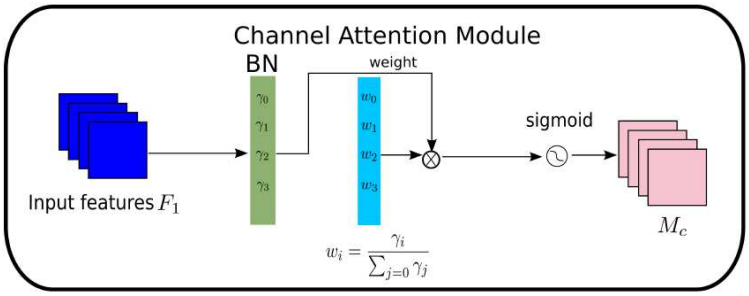
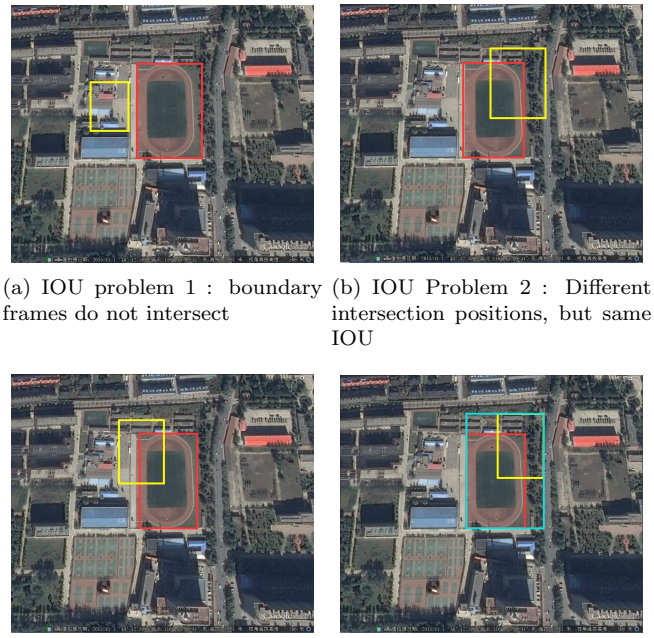


Figure 5: NAM Space Attention Module

3.5 Generalized cross point joint loss

IOU is the most commonly used evaluation index in target detection, and it is also commonly used as boundary box regression loss. However, when the prediction box and the real box do not cross directly (Figure 6 (a)), IOU loss cannot be gradient optimized. In addition, IOU can ' t distinguish alignment in different ways, there is the same IOU value when the prediction box and the real box cross in different directions (Figure 6 (b) and (c)). In order to solve such problems, Hamid Rezatofighi et al. proposed the generalized joint intersection loss (GIOU Loss)[32]. On the basis of IOU, GIOU considers the outer rectangle between the two boxes (Figure 6 (d)), which can better distinguish the relative position of the boundary box and accurately realize the regression of the boundary box.



(a) IOU problem 1 : boundary frames do not intersect

(b) IOU Problem 2 : Different intersection positions, but same IOU

(c) IOU Problem 2 : Different intersection positions, but same IOU

(d) GIOU calculation

Figure 6: CSP module structure diagram (left graph is YOLOX, right graph is SNG-YOLOX)

4 Experiment

4.1 Experimental Environment and Model Parameters

This experiment uses the SNG-YOLOX-S model implemented by Pytorch 1.11 and performs training and testing on NVIDIA RTX3080 GPU, 12th Gen Intel (R) Core (TM) i7-12700KF CPU and 32 GB memory.

The epochs of SNG-YOLOX-S model are set to 200 rounds, which are trained twice, 100 rounds each time. For each 100 rounds, the batch size of the first 50 rounds is set to 8, the learning rate is set to 4, and the learning rate is set to 4. The learning rate adopts exponential decay strategy, the optimizer adopts Adam optimizer, and the data set adopts MOSAIC data enhancement.

4.2 Introduction of data sets

This paper uses RSOD remote sensing data set for experiments. RSOD dataset is a remote sensing dataset launched by Wuhan University in 2015, which contains 976 images and 6950 columns. The dataset contains 4993 aircrafts, 191

playgrounds, 180 overpasses and 1586 oil tanks with spatial resolution ranging from 0.3m to 3m. Some images in the data set are shown in Fig. 7, and the results of the number of various types of images and the number of detection boxes in each column in the statistical data set are shown in Fig. 8.

Figure 7: RSOD dataset image

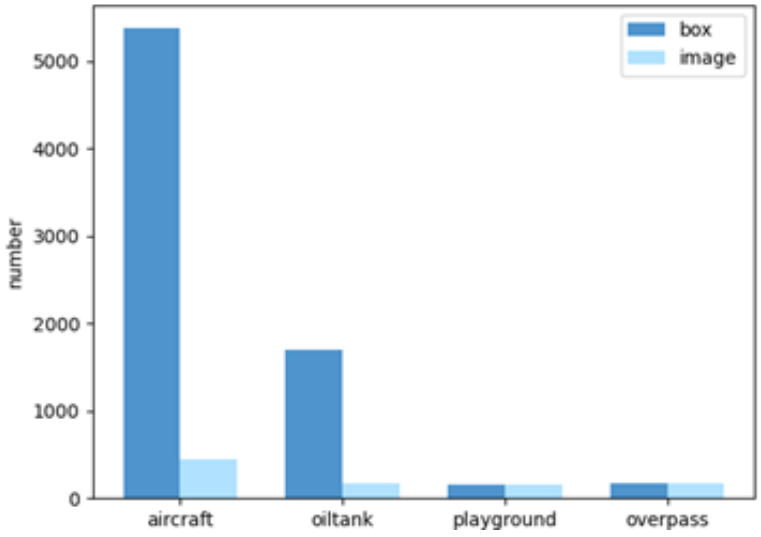


Figure 8: RSOD dataset image and real column number statistics

It can be seen from Figure 7 that the difference between the target and the background in the overpass image is small, and there is no obvious boundary, which may affect the accuracy of the model prediction. In addition, the target in the playground image has obvious angle problem, and the data set uses a rectangular annotation box rather than a rotating annotation box, so it is bound

to bring deviation of detection accuracy in actual prediction. It can be seen from Fig. 8 that there is an obvious problem of uneven data distribution in the RSOD data set. The data of aircraft and fuel tank are relatively sufficient, while the data of Cao Cao and overpass are obviously insufficient. Direct training is bound to cause poor detection performance of the model on some samples.

In general, in the process of model construction, the influence of small difference between detection target and background and uneven sample on model detection performance should be considered, and appropriate feature extraction methods and uneven sample processing methods should be selected. Therefore, this paper introduces the NAM module to enhance the attention of non-indigenous regions of the image, and uses MOSAIC to reduce the influence of target size difference and sample unevenness on the results.

4.3 Experimental result

4.3.1 Target detection results

SNG-YOLOX-S algorithm was used for experiments on the RSOD data set, and the obtained detection results are shown in Fig. 9.

Figure 9: Partial image detection results of RSOD dataset

As can be seen from Figure 8, as in the previous data set analysis, the data set has the same impact on the model detection. In the real detection of the experiment, the model has better detection effect on the images of aircraft and fuel tank with more data and examples, while the detection effect on the overpass and playground images is poor. Among them, the model can still achieve better detection performance on smaller targets (such as Figure 9 aircraft image) ; for overpass images, due to the small difference between overpass images and background and no obvious boundary, the model tends to predict a larger detection box, resulting in loss of accuracy ; because of the rotation of the playground image, and the actual marker box is not a rotating marker box, there is

a certain gap between the detection performance of the model and the aircraft and the barrel.

4.3.2 Algorithm comparison

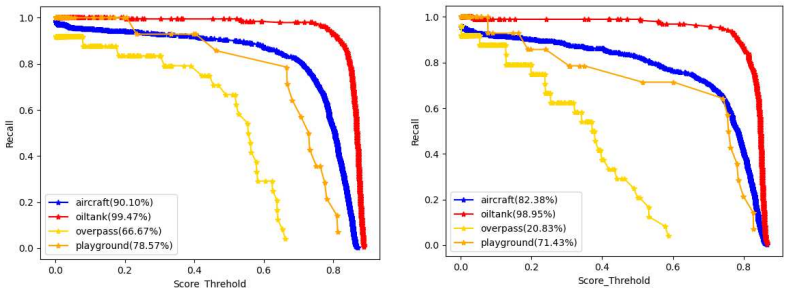
In order to compare the detection performance of this algorithm with other target detection algorithms on remote sensing data sets, this experiment selects YOLOv4[12], YOLOv4-TINY[34], YOLOv3-EFFICIENTNET[35], YOLOv5-S[8] and YOLOX-S[7] algorithms to compare the average detection accuracy and detection accuracy on various categories with SNG-YOLOX-S algorithm in this paper. The results are shown in table 1.

Table 1: Comparison results of detection performance of various algorithms

Methods	mAP(%)	aircraft	oiltank	palyground	overpass
YOLOv4	60.21	83.61	94.06	57.99	5.17
YOLOv4-TINY	60.21	57.46	95.24	72.95	15.20
YOLOv3-EFFICIENTNET	70.52	74.33	92.92	89.75	25.10
YOLOv5-S	57.46	84.73	98.15	42.02	4.94
YOLOX-S	83.62	91.58	99.62	91.83	51.44
SNG-YOLOX-S	**91.47**	**95.27**	**99.71**	**94.33**	**76.55**

It can be seen from Table 1 that the SNG-YOLOX-S algorithm has achieved advanced detection performance in the overall average accuracy and the detection accuracy of each category. After the improvement of the original YOLOX-S algorithm combined with the idea of receptive field, the attention of non-indigenous features and the loss of boundary box regression, the performance of the original YOLOX-S algorithm can be improved slightly on the target box and the oil tank image with more data volume (0.09 % mAP). A small amount of improvement (3.69 % mAP) can be achieved on the aircraft data set containing more small targets, a small amount of improvement (2.5 % mAP) can be achieved on the large target playground image, and a substantial improvement (25.11 % mAP) can be achieved on the non-indigenous overpass image. The reason is that the introduction of self-calibration convolution (SCNet) enhances the model ' s perception of image context, and the introduction of normalized attention mechanism (NAM) module can better focus on the non-indigenous features in the original image (small targets and targets with small background and target discrimination), which improves the performance of the model. The introduction of GIOU loss is conducive to better regression of boundary frames, so it also brings performance improvement on the basis of other algorithms. In contrast, other target detection algorithms can achieve better detection performance when the images to be detected and the target frame are sufficient (such as the barrel image), but the detection performance is poor when the small target and the target are not obvious.

In order to better compare the SNG-YOLOX-S algorithm with YOLOX-S algorithm, this paper analyzes it from the perspective of recall rate, as shown in Figure 11.



(a) SNG-YOLOX-S (b) YOLOX-S

Figure 10: Algorithm precision - recall curve

It can be seen from Fig. 10 that SNG-YOLOX-S algorithm has achieved priority detection performance in terms of average accuracy and recall rate in each category. Among them, the recall rate of SNG-YOLOX-S algorithm is higher than that of YOLOX-S algorithm in the overpass image with no obvious target and background, reaching 45.84 %, which is greatly improved. The recall rate in other categories has been improved to a certain extent, which is consistent with the analysis results of the model in the average accuracy above.

4.4 Ablation experiment

In order to analyze the specific effect of each module, the experiment is carried out by policy stack, and the results are shown in table 2.

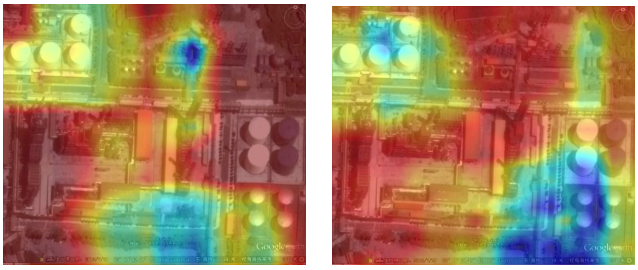
Table 2: Experimental analysis of SNG-YOLOX-S ablation

Methods	mAP(%)	aircraft	oiltank	palyground	overpass
YOLOX-S	83.62	91.58	99.62	91.83	51.44
YOLOX-S+SCNet	88.41(↑4.79)	93.35	99.49	87.31	73.47
YOLOX-S+SCNet+NAM	91.03(↑2.62)	95.13	99.82	92.29	76.89
YOLOX-S+SCNet+NAM+GIOU	91.47(↑0.44)	95.27	99.71	94.33	76.55

It can be seen from Table 2 that after SCNet was used to replace the deep convolution module in the original CSPLayer on the original YOLOX-S, the overall performance of the model was improved by 4.79 % mAP, which greatly improved the detection performance of the original model on the non-indigenous overpass target, up to 22.03 % AP, and also improved 1.77 % AP on the aircraft

image containing more small targets. The introduction of SCNet also brings a little performance degradation in oil buckets and playground images. The reason is that the introduction of SCNet takes too much information of the surrounding context into account, resulting in interference with local information. Therefore, this paper continues to introduce the NAM module into the SCNet output terminal on the basis of SCNet to enhance the attention of local and non-obvious information. After the introduction of NAM, it can be found that the detection accuracy of the model in each category has been further improved, which exceeds the original algorithm, and the overall detection performance is improved by 7.41 % mAP compared with the original algorithm. Considering the defects of IOU in the boundary frame regression, the GIOU loss is introduced for the boundary frame regression. Finally, compared with the SCNet and NAM strategies, the improvement of 0.44 % mAP is also brought, especially in the playground image of large targets, reaching 2.04 % AP. However, due to the divergence of GIOU in the training process, it also causes a slight decline in detection accuracy in other categories. In addition, it is found in this experiment that DIOU and CIOU are more stable than IOU and GIOU, and are more consistent with the mechanism of target box regression. However, it is found in this experiment that DIOU and CIOU are less effective than GIOU on this dataset. The reason is that although DIOU[36] and CIOU[37] can carry out the regression of the target frame more accurately, due to the obvious rotation of some images (such as the playground image) in the RSOD data set, there is a difference between the marker frame and the real position, and some images have no obvious boundary (such as the overpass image). There is a contradiction between the more accurate boundary frame regression and the actual position of the target learned by the model, resulting in the lack of performance of these two loss functions used in this data set.

In order to analyze the difference between SNG-YOLOX-S and YOLOX-S more intuitively, the effect of introducing modules is analyzed. Here, Grad-CAM++ is used to visually explain the CNN model prediction of the output part of the trunk network (CSPDarknet), so as to analyze the effect of the introduced module on target localization. Among them, GradCAM++[38] is better than GradCAM [39] in explaining multiple object instances in a single image. The results are shown in Figure 11.



(a) SNG-YOLOX-S thermal diagram (b) YOLOX-S thermal diagram

(c) SNG-YOLOX-S detection results (d) YOLOX-S detection results

Figure 11: Analysis of Model Feeling Field

From the perceptual field visualization results of the two algorithms on the barrel image in Figure 11, it can be seen that the perceptual range of SNG-YOLOX-S algorithm (Figure 11 (a)) is larger than that of YOLOX-S algorithm and the perceptual region (Figure 11 (b)) is more accurate (focusing on the position of the barrel). It can also be seen from the detection results of the two algorithms that the detection accuracy of SNG-YOLOX-S algorithm is higher (Fig. 11 (c)), and the detection accuracy of YOLOX-S is relatively low (Fig. 11 (d)), and there is a misjudgment of the detection target in the upper left corner of the figure. It is found that the YOLOX-S algorithm misjudges the position in (d), and the algorithm in (b) has relatively weak feeling intensity in this region.

In summary, the introduction of SCNet, NAM and GIOU on the basis of YOLOX-S in this paper can enhance the perception range of the original model, improve the extraction ability of the model on non-dominant features, and have better detection performance.

5 Conclusions

In this paper, in order to improve the detection performance of target detection algorithm on non-obvious remote sensing image targets. Based on the YOLOX-S model, we add some advanced technologies, namely self-calibration convolution module (SCNet), normalized attention module (NAM) and

generalized intersection joint loss function (GIOU Loss), and named SNG-YOLOX-S. Using SNG-YOLOX-S to perform experiments on RSOD remote sensing data sets, it is found that SNG-YOLOX-S has higher detection performance for non-obvious targets. Compared with other target detection algorithms, SNG-YOLOX-S algorithm has achieved priority detection performance. We hope that this report can help developers and researchers to provide reference for model design in the detection of non-obvious remote sensing targets.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.61966022), the Natural Science Foundation of Gansu Province (No.21JR7RA300) and the Open Project of Dunhuang Cultural Heritage Protection Research Center of Gansu Province (o.Gdw2021Yb15).

7 Data Availability Statement

The data generated during this study are included in the article and its supplementary information.

References

[1] Wang Xili, Liang Min, Liu Tao. Feature-enhanced single-stage remote sensing image target detection model [J/OL]. Journal of Xi 'an University of Electronic Science and Technology : 1-11 [2022-04-13].http://kns.cnki.net/kcms/detail/61.1076.tn.20211123.1524.006.html

[2] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018

[3] Dong Ruchan, Jiao Li Cheng, Zhao Jin, Shen Weiyan. A deep fusion mechanism of remote sensing image target detection technology [J].Journal of Xi ' an University of Electronic Science and Technology, 2021,48 (05) : 128-138.DOI : 10.19665 / j.issn1001-2400.2021.05.016.

[4] Zhao Yongqiang, Rao Yuan, Dong Shipeng, Zhang Junyi. Overview of deep learning target detection methods [J]. Chinese Journal of Image Graphics, 2020, 25 (04) : 629 – 654.

[5] Zhu Y, Zhao C, Wang J, et al. Coupling global structure with local parts for object detection[C]//Proceedings of the IEEE international conference on computer vision. 4126-4134.

[6] Jiaojiao, Hu Yongli, Sun Yanfeng, Yin Baocai. Summary of small target detection methods based on deep learning [J].Journal of Beijing University of Technology, 2021,47 (03) : 293-302.

[7] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[8] glenn jocher et al. yolov5. https://github.com/ ultralytics/yolov5, 2021. 1, 2, 3, 5, 6

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In CVPR, 2016. 1

[10] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In CVPR, 2017. 1, 3

[11] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. 1, 2, 3

[12] Alexey Bochkovskiy, Chien-Yao Wang, and HongYuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020. 1, 2, 3, 6

[13] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 390–391, 2020.

[14] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, pages 6105–6114. PMLR, 2019.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021.

[16] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.

[17] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2778-2788.

[18] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[20] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In CVPR, pages 1–9, 2015. 2

[21] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In CVPR, pages 2403–2412, 2018. 2

[22] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and ´ Kaiming He. Aggregated residual transformations for deep neural networks. In CVPR, pages 5987–5995, 2017. 1, 2, 4, 5, 7

[23] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016. 1, 2

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, pages 7132–7141, 2018. 1, 2, 4, 5, 7

[25] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In NeurIPS, pages 9401–9411, 2018. 1, 2, 4, 7

[26] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.

[27] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. arXiv preprint arXiv:1904.11492, 2019. 1, 2, 7

[28] Liu J J, Hou Q, Cheng M M, et al. Improving convolutional networks with self-calibrated convolutions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10096-10105.

[29] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514, 2018

[30] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3139–3148, 2021.

[31] Liu Y, Shao Z, Teng Y, et al. NAM: Normalization-based Attention Module[J]. arXiv preprint arXiv:2111.12419, 2021.

[32] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 658-666.

[33] Long Y, Gong Y, Xiao Z, et al. Accurate object localization in remote sensing images based on convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2486-2498.

[34] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-yolov4: Scaling cross stage partial network[C]//Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. 2021: 13029-13038.

[35] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[36] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12993-13000.

[37] Zheng Z, Wang P, Ren D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021.

[38] Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. arXiv 2017[J]. arXiv preprint arXiv:1710.11063.

[39] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

[40] Li Z, Peng C, Yu G, et al. Detnet: A backbone network for object detection[J]. arXiv preprint arXiv:1804.06215, 2018.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SNGYOLOXS.rar
- dataSet.rar